Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials

A project submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Professional Studies

Betty Anne Schwarz

Student Number: **M00460456**
Project Module: DPS5360

May 16, 2016

Word Count: 59,158

Contents

List of Tables:

List of Figures:

Abstract

Summary

Within the health sciences, action research is a methodology well suited to the goal of collaboratively improving practice. As the Royal College of Radiology recommends the use of published clinical trials as guides for achieving higher standards of accuracy, it is important for radiologists to reflect deeply on the results from diagnostic accuracy studies. When the results of the gold standard (or reference standard) are used to confirm a particular diagnosis or disease by comparing the diagnostic accuracy to a newer or index test, this is referred to as diagnostic accuracy research. In the reporting of all research, every effort must be made to reduce the incidence of bias. In 2003, the STARD (Standards for Reporting Diagnostic Accuracy) tool was developed for clinicians to enhance the quality of reporting diagnostic accuracy studies. Based on previous studies, experiential knowledge, and an extensive review of the literature, this research demonstrates that the STARD tool is not being fully optimized. The overall aim of this research was to conduct a work-based project within the department of radiology to develop a revised tool, based on the current STARD, which could then be used to more accurately report and interpret the results of radiology diagnostic accuracy trials. This study was conducted in accordance with participatory action research.

Methods

The development of this new reporting tool was conducted in collaboration with a group of physicians, and in two distinct phases. First, a needs assessment was sent to eight radiological experts who had agreed to participate in the study. Based on their responses, and feedback from my mentor and colleagues, the next phase of tool development was done using the Delphi technique, after two rounds of which consensus was met. Each phase and cycle iteration to complete the needs assessment and Delphi technique are synonymous with the cycles of action research. The new reporting tool was named the RadSTARD (Radiology Standards for the Reporting of Diagnostic Accuracy Studies), and an elaboration document was written to provide guidance to the end-user. Radiology residents and Fellows at The Ottawa Hospital were then asked to rate their level of confidence in interpreting a diagnostic accuracy article specific to radiology while referring to the RadSTARD. They were also provided a second diagnostic article, the STARD tool, and an elaboration document for comparison. Data was collected using questionnaires that allowed for additional comments.

Findings

The validation phase of the RadSTARD tool was completed via triangulation of data, as both a quantitative and qualitative analysis was completed. The results found no significant statistical difference between the two groups as per the Mann-Whitney and chi-square analysis. Likewise, both physician

groups indicated that they found RadSTARD increased their level of confidence when interpreting the diagnostic accuracy article. Concomitantly, when combined, 96% of the two physician groups indicated they would use the tool again.

Interpretation

These results may be interpreted as generalizable, as there was no discrepancy or statistical difference found in the results between the radiology residents' and Fellows' scores, despite the differences in their level of training. Both groups found the RadSTARD tool and elaboration document to be beneficial to them when interpreting the literature. RadSTARD is thus a reliable tool that can be used to validate the results of diagnostic accuracy studies specific to radiology. It will aid radiologists in reporting and interpreting radiology diagnostic accuracy studies, impacting their practice for generations to come.

Glossary:

Action research cycle: four essential steps required between each iteration of self-reflective spirals conducted according to the goals and purpose of the research study. These steps include planning, action, observation and reflection

Acute hematogenous osteomyelitis: an infection of the bone that affects young children

α error: the probability of rejecting a true null hypothesis

Bias: over-interpretation of results

Blinded interpreter: when each investigator providing an interpretation is unaware of another's investigator's interpretation

Bone scintigraphy: a nuclear medicine that determines severity of bone fractures or disease via the use of radioactive material
BMJ: British Medical Journal

Chi-Square (or Fisher exact tests): assess the proportions expressing confidence for each question from the raw data

Cochrane: an independent network of researchers who work together to provide reliable evidence based reviews of the literature known as Cochrane reviews

Chondrosarcomas: cancer of the cartilage

Confidence intervals: a statistical measure that describes the degree of certainty within a particular sample

CONSORT: Consolidated Standards for the Reporting of Trials

CMR: cardiac magnetic resonance imaging

Critical friend: facilitator to action researchers as consultants and provide academic guidance

CT: computerized tomography

3D motion measurement: define the location of items in 3-D space

Diagnostic accuracy: how well a test can diagnose a condition or rule it out

Delphi technique: a method of collecting information from a panel of experts in a systematic fashion. Questionnaires are typically employed for 2 rounds or more of the Delphi technique until consensus has been met

Double-loop-learning: questioning of underlying governing variables by subjecting them to critical scrutiny

DXA: dual-energy X-ray absorptiometry

EMBASE medical literature database

Endometriosis: when endometrial tissue resides outside the uterus and results in pelvic pain

Epistemology: current knowledge about a particular area of study

EQUATOR Enhancing the QUAlity and Transparency Of health Research)

Evidence of spin: over-interpreting their trial results

Generalizability: the determinate of reliability of measurements that are under study

ICMJE: International Committee of Medical Journal Editors

Index test: the test under study

JAMA: Journal of American Medicine

Kappa statistic: is the degree of measurement of categorical items

Likert rating scale: a rating scale that is used by respondents to provide their responses along a particular range

Liver-Fibro-STARD: compared to STARD, additional items added pertaining to diagnostic accuracy studies on liver fibrosis tests

Mann-Whitney Test: ranking individual scores of a given sample and then comparing the results between two groups

MeSH: Medical Subject Headings

MEDLINE: medical literature database

Menorrhagia: abnormal heavy menstrual bleeding

MRI: magnetic resonance imaging

Needs assessment: a way of identifying needs or gaps between current knowledge and need or want to fill that gap

NEJM: New England Journal of Medicine

Ontology: the subject of existence. It is the concepts that already exist for an individual or a community as a whole.

Participatory action research: a research method that emphasizes participation amongst collaborators and action via a collective inquiry

PET: positron emission tomography

PRISMA: Preferred Reporting Items for Systematic Reviews

p-value: is a statistical value that expresses the results obtained from hypothesis testing

QUADAS: Quality Assessment of Diagnostic Accuracy

QUADAS-2: Quality Assessment of Diagnostic Accuracy - Revised
Qualitative data collection methods: conducting interviews and observation

Quantitative data collection methods: statistical analysis

QUOROM: Quality of Reporting of Meta-analyses

RADART: Radiology Diagnostic Accuracy Reporting Tool

RadSTARD: Radiology Standards for Reporting Diagnostic Accuracy Studies

RCR: Royal College of Radiology

Reference standard: gold standard

Review bias: lack of information provided to those interpreting the test results

ROC: Receiver Operator Characteristic – is a graphical plot that is created to plot the true positive rate (TPR) against the false positive rate (FPR) for diagnostic decision making

Sensitivity and specificity: are the statistical measures used to determine true positive (sensitivity) from true negative (specificity) for patients either having or not having a particular condition under study

Signal-to-noise ratio: is a measure used in radiology that affects image interpretation which compares signal levels to background noise

Single loop learning: whereby the loops of action research link to the reflection involved within the framework of an organization

SORT: Strength of Recommendations Taxonomy

SPECT: single positron emission computed tomography
STARD: Standards for the Reporting of Diagnostic Accuracy Trials

STARDdem: Standards for Reporting of Diagnostic Accuracy – Dementia

Systematic review: analysis of multiple studies

STARD- DI: Standards for Reporting of Diagnostic Accuracy - Diagnostic Imaging

Tetralogy of Fallot: congenital heart defect

TiDier: Template for Intervention Description and Replication

Transdisciplinary: research conducted by investigators collaborating together in a multi-disciplinary approach

Transdisciplinarity: is a research standard that was originally developed to deal with research issues pertaining to societal needs

Acknowledgements

My doctoral journey began several years ago with a soft knock on my office door by a young physician who out of the blue came to encourage me to embark on my Doctorate. His name was Dr. Adnan Choudhry. He came with a mission and that was to strongly encourage me to complete my doctorate after he learned that I had received the Student of the Year Award from the University of Liverpool for my MSc. Although my department congratulated me on many occasions I'll always remember how Adnan repeatedly smiled at me and said, "You just have to do it". Truthfully for years, I had strongly contemplated advancing on to doctoral studies after I completed my masters, but somehow this visit from a physician who barely knew me on that sunny morning spurred me on. During the ensuing months, I scoured the web in pursuit of a doctoral degree that would incorporate my work with my studies. It was also important to me to find a doctoral degree that would allow me to follow-up to the results from my master's thesis. Hence, my first contact with Middlesex University was made when I called Dr. Ali Mehmet Dickerdem to discuss my proposal for my doctorate. He was quite enthused and as they say, 'the rest is history'. Thank you very much Dr. Ali Dickerdem for taking that initial call from me and for your support then and all throughout my doctoral program.

Over the years, I have been very fortunate to have been supported by Dr. Wael Shabana. Dr. Shabana is a radiologist who also holds a PhD and has been there to advise me first on my masters and again as my mentor for my doctorate studies. He is quite versed in research methods and statistics. I can't thank him enough....

Once I began my studies, I worked closely with Dr. Wael Shabana who was my academic consultant and initially Dr. Barbara Workman who was my academic advisor from Middlesex University. Thank you, Dr. Workman for guiding me through the initial phase of my doctorate. When Dr. Workman retired, Dr. Annette Fillery-Travis was assigned to me. I truly marvelled at her knowledge and ability to understand where I was coming from and where I wanted to go! Thank you so much Annette for being there to guide me when I had clearly gone off course. I learned a great deal from you and I will always be grateful for the on-going support and tutelage provided to me throughout this challenging degree program. I was extremely honored to receive the Santander Scholarship and hence I travelled to Middlesex University to receive this prestigious award. While I was there I was very blessed to meet Dr. Annette Fillery-Travis, Dr. Paul Gibbs, Dr. Kate McGuire and Dr. Mehmet Ali-Dickerdem in person. Thank you for your gracious welcome and encouragement and I can't wait to meet you all again. I was also moved by the words of encouragement extended to me by the Chancellor, Dame Jane Ritterman. Thank you as well to Stephen

Watt, Senior Program Administrator of the program for his endless guidance and support and the librarians from the university library.

As my study consisted of an action research study, I was fortunate to be in receipt of additional guidance from my critical friends, Dr. Swapnil Hiremath and Dr. Nitin Ramamurthy. Their advice was greatly appreciated. In addition, I was most fortunate that Dr. David Moher was on staff at The Ottawa Hospital. Dr. Moher is a world renowned scientific research methodologist who specializes in reporting tools. I had previously met Dr. Moher many years ago during my professional career and I was so ever grateful when he agreed to meet with me to discuss my methodology. David provided valuable advice and suggested I contact Dr. Jennifer Hoffman. Dr. Hoffman graciously provided me support based on a new reporting tool she had recently developed called the TiDIER checklist.

Conduct of this study required a very extensive review of the literature which was completed three times by the very talented Sascha Davis who is a library scientist from The Ottawa Hospital. Many thanks to Sascha for always being there to help me - your kindness will never be forgotten. I would also like to acknowledge the rest of the library staff who helped me retrieve any articles that were not available from the Middlesex University library.

From the department of radiology, I have received support from nearly the entire department. I am most thankful. Most notably; Dr. Mark Schweitzer initially met with me to discuss my idea for my doctorate and provided valuable additional insight. Many thanks to Dr. Schweitzer for that initial advice. I would also like to thank Dr. Schweitzer and Mario Villemaire for the sponsorship support from the department to complete my doctorate. I would also like to thank Dr. Jose Aquino for his support when he was Interim Chief. Most recently, Dr. Frank Rybicki – our new Chief rendered his support for the validation phase of my doctorate. I was very grateful for that support as we didn’t really get an opportunity to discuss my study in great detail yet, he believed in me and the concept. Thank you Dr. Rybicki.

I am very thankful to the radiologists who accepted my invite to participate in my study. These radiologists were the radiological experts who collaborated with me to develop the RadSTARD tool. They include Dr. Carole Dennie, Dr. Carolina Souza, Dr. Andrew Scott-Montcrieff, Dr. Santanu Chakraborty, Dr. Nicola Schieda, Dr. Rebecca Thornhill, Dr. Marcos Sampaio and Dr. Ashish Gupta. Validation of the newly developed RadSTARD tool was done by the radiology residents and Fellows. I am most thankful to them and in particular, Dr. Mike Connolly (Chief Resident) and the Academic Heads for the residents and Fellows – Dr. Rebecca Hibbert and Dr. John Ryan. Thank you as well to William Petricich who calculated my statistics and proof read my interpretation of the stats. Your statistical expertise is in great demand for the physician group, yet you always found time to help me. I was very humbled by your kindness William – many thanks.

I am also most thankful for the assistance of Kelly Hawkes from the OHSN-REB (Ottawa Health Sciences Network - Research Ethics Board), Kathy Millar and Dr. Raphael Saginur (Research Chair) for their guidance and ethics approvals after many renditions of my study documents.

On a personal level, the love and support of my family throughout the years of my academic pursuits will never be forgotten. My mother listened to me on countless occasions as I described my thoughts or my next steps of my doctoral studies. She used to love it when I read some of my chapters to her as she was a teacher. When I was done she lovingly encouraged me with the words, "I live again through you". I know her endless prayers gave me the strength to never stop. Thank you so much Mom – I love you dearly. My Dad also supported me and marvelled at my determination. He always knew I could take on this monumental amount of work in addition to my full time job. I love you Dad for believing me. A big thank you to my brother Tony, who also encouraged me on so many occasions with strength in his eyes as he said, "believe in yourself". His belief in me gave me the fortitude to carry on. As my dissertation was coming to the end, I will never forget him saying, "you remind me of the time when we ran relay races in high-school. You have the final baton in your hand and you're gaining speed as you run toward the finish line". Love you Tony! My sister Lynda, her family and my younger brother Pat also supported me throughout the years and sent their love across the miles. Thank you so much.

Last but not least is my husband Klaus who at the beginning of this academic journey I will never forget the sentiment he wrote in a card he gave me that read inside, "I can't wait to call you Dr. Schwarz my darling". He has always believed in me and provided me endless love and support throughout. Therefore, my dear Klaus I dedicate my doctorate to you for you have certainly been there for me along every single step of my academic pursuits.

Finally, I need to mention my four-legged furry son, Schwarzie who silently supported me every day as he lay beside me on his bed or the floor, often sleeping mind you as I read and typed into the wee hours of the morning on so many occasions. Thank you Schwarzie for simply just being there! And to all my friends and extended family who have also supported me and encouraged me all throughout my doctorate – thank you all so very much. I did it!

1. Chapter 1 Introduction

1.1 Planning my doctoral study

For over 20 years now I have worked in clinical research. Initially, I worked in clinical trials as a research coordinator. Typically these positions require a registered nurse, as there are many clinical assessments to conduct. In the late 1980s, I was working in a hybrid position as the Medical Team Coordinator within the division of neurology at The Ottawa Hospital. My mandate was to improve the efficiency of the medical team, which meant I worked very closely with all members. Due to my overwhelming interest in learning more, I routinely attended academic conferences and became so interested in the neurosciences that I also started attending weekend lectures with the physicians who were preparing for their neurology Royal College exams.

Several years later, I was introduced to clinical trials. Within this academic milieu of extended learning my knowledge had grown exponentially, as I learned how to perform physical and neurological exams. As such, when the clinicians (who were also Co-investigators) were planning an Alzheimer's trial I was called upon to admit the study participants. I also took first call (via a pager) should the participants remove their cardiac monitors that were measuring the incidence of bradycardia that could occur from the drug that we were testing. Although I wasn't a doctor, the physicians knew about my additional informal training, so I took on the role of clinician for the Alzheimer's trial. I knew then that somehow I had entered into an exciting pedagogical environment based on my experiential work-based learning, supplemented by self-study (Brodie and Irving, 2007).

As I reflect upon my previous experiences and education throughout my professional career, I realize how fortunate I was because they prompted me to continually advance my knowledge (Grix, 2004). To this day, I attribute these key milestones to my professional and epistemological learning development. Advancing my learning as my career evolved, either experientially or formally, has also been a dominant pattern that has impacted my ontological development (Koshy et al., 2011). For example, once I was exposed to neuroscience medical rounds and academic lectures within the department of neurology, I was inspired to extend my formal learning and obtain my undergraduate degree in 1991.

Throughout the ensuing years my research and clinical responsibilities continued to become more complex as I took on more advanced responsibilities. Due to major restructuring within our hospital infrastructure, the neurology ward closed in 1998. I decided to embark upon other positions within the neurosciences, which further enhanced my ontological development. Having branched into Parkinson's disease as the clinic nurse, I was later offered the opportunity to learn how to program deep brain stimulators. This was a very novel approach of treating patients with intractable tremors and abnormal movements, and I remember how excited I was to take on this challenge. We successfully ran this multidisciplinary approach, combining neurosurgery and neurology, to develop treatment strategies for our surgical candidates for several years, until the surgeon left. I became discouraged and decided to embark on new career challenges outside of the hospital. Although I found these positions less satisfying, I learned that no matter which career path I chose, it further enhanced my epistemological development, as I was always seeking to learn more. On further reflection, I now appreciate how these opportunities allowed me to advance certain skills that I didn't realize would be so necessary until now.

Several years later, in 2007, I returned to the hospital for a position as a research coordinator in radiology. At that time, I was working mainly in neuroradiology research, which was relevant given my background. I considered this career move as the most significant milestone in my professional epistemological development, as this position motivated me to complete my graduate degree in research. Throughout this transition, as I advanced my knowledge and career in clinical trials, I simultaneously found that my level of scientific inquiry had also grown. As our radiology research program grew, so too did my responsibilities – so much so that I was provided a full-time research assistant and promoted to Clinical Research Program Manager.

At this time, I was finalizing my Master's degree in clinical research. I tried at every opportunity to use my knowledge in my work within the department of radiology. For example, I began writing research protocols and developed case report forms for some of radiologists and trainees. In this role, I also provided departmental presentations on what the requirements were for conducting research, whether prospective or retrospective in nature. The final stage of my Master's degree required an independent study and write-up of the findings in a dissertation. I wanted to study a topic that could perhaps benefit the department. The radiological sciences are dynamic, as new technology is constantly being developed. When the results of the gold standard (or reference standard) are used to confirm a particular diagnoses or disease entity by comparing the diagnostic accuracy to a newer or index test, this is referred to as diagnostic accuracy research. In the reporting of all research, every effort must be made to reduce the incidence of bias (Bossuyt et al., 2003a). This can be achieved with the use of reporting tools (Moher et al., 2014b).

Given my position in radiology research, I chose to conduct a systematic literature review for my Master's thesis to determine if there had been an improvement in the quality of reporting and methodology of diagnostic accuracy trials since the STARD (Standards for the Reporting of Diagnostic Accuracy Trials) and QUADAS (Quality Assessment of Diagnostic Accuracy) were developed in 2003. In keeping within the domain of radiology, I based my review of diagnostic accuracy studies on patients presenting with acute abdominal pain to the emergency department. I concluded that there was evidence of overall adherence to the guidelines. However, given that these tools are generic, I decided that perhaps a revision to the STARD would make it much more conducive to reporting results, and thus diagnostic radiology studies would be more accurate and more beneficial to radiologists (Bossuyt et al., 2003a, Whiting et al., 2003).

Although our department is robust with research activity, of those that are considered diagnostic accuracy trials, the current reporting tools such as the STARD are not employed to their full capacity. This finding is elaborated upon in my literature review chapter. I attribute this lack to the limitations of the currently available tools. As I do see the value of their use, my doctoral research will focus on developing an evidenced-based tool that can be used to report and interpret radiology diagnostic accuracy trials. I consider my ambition to continue my post-graduate studies as the pinnacle of my epistemological development of scientific inquiry and discovery as I introduce my proposed research for my Doctorate of Professional Studies.

Thus, the overall aim of my doctoral research is to conduct a work-based project within the department of radiology to develop a revised tool, based on the current STARD, which could be used to accurately report and interpret the results of radiology diagnostic accuracy trials. Given my position within the division of radiology, and my close working relationship with those involved in research, my project proposal consisted of an action research study whereby I worked with the radiologists to first determine if they thought there was a need to develop a revised tool specific to radiology – the RadSTARD (Radiology Diagnostic Accuracy Reporting Tool). Providing this was the case, once the new and/or revised tool was developed, an elaboration document was also developed which describes each item of the newly revised tool with radiology examples that illustrate how such a tool may benefit radiologists and their trainees. Then I tested the revised tool with radiology residents and Fellows to validate the merits of the tool specific to radiology diagnostic accuracy trials.

Based on this brief introduction to my proposed methods, I chose action research due to the participatory nature of cyclical research, flexibility, and reflective properties to conduct my doctoral study within our radiology department. In particular, I utilized participatory action research, as it is a social process that is participatory, practical and collaborative, emancipatory, critical, and reflexive with an overall goal of changing both theory and practice (Kemmis and McTaggart, 2005).

1.1.1 Achieving Standards of Accuracy in Radiology

As the Royal College of Radiology recommends the use of published clinical trials as a guide in achieving standards of accuracy, it is important for radiologists to reflect upon the results from diagnostic accuracy studies (RCR, 2012). This will decrease the use of inappropriate investigations, resulting in enhanced patient care as the proper diagnostic test is chosen based on the results from the accurate reporting of the sensitivity and specificity analysis of diagnostic accuracy trials. Within healthcare, the purpose of testing is to provide a constellation of information that is necessary to the decision-making processes. In order to comprehend the results of each test, it is essential to understand the multiple purposes of a single test; indeed, the same test can be used in numerous situations depending on its overall purpose. Regardless of which test is chosen, the term “index test” refers to the test results that are being evaluated (Riegelman, 2000). As diagnostic radiology is influenced by the many evolving technologies described within the literature, it is important that the results of the findings are accurately interpreted, allowing for appropriate use in clinical applications (Jones and Athanasiou, 2009).

1.1.2 Diagnostic Accuracy Studies

In diagnostic accuracy studies, the results of the index test are evaluated in terms of the definitive gold standard or reference standard (Riegelman, 2000). The nature of the patient's clinical condition or clinically suspicious diagnosis will prompt the clinician to use the reference standard that is considered most relevant to confirm the diagnosis. Consequently, the reference standard could be either a laboratory or diagnostic imaging test, or it could be a combination of tests that are utilized in conjunction with routine follow-up care for the patient. Accuracy is determined by the level of agreement within the index test or test under evaluation, as it is compared to the reference standard. Therefore, diagnostic accuracy can be illustrated in a number of ways, including sensitivity and specificity, ROC (receiver operator characteristic) curves, likelihood, and odds ratios (Bossuyt et al., 2003b). Diagnostic accuracy studies compare the results of the index tests to the reference standard, within a consecutive series of patients (Manchikanti et al., 2009). The results of diagnostic accuracy studies are considered vital in the assessment of new or existing diagnostic tools. Given that the results of diagnostic accuracy studies are often used to guide future patient care, it is paramount that the quality of reporting of the trial results is done completely and accurately. The benefit of accurately reporting trial results are twofold: it enables the reader to assess for potential bias while simultaneously evaluating the overall generalizability of the study results (Bossuyt et al., 2003b).

1.2 The STARD (Standards for Reporting of Diagnostic Accuracy) Statement

The STARD statement was developed in 2003 by a group of scientists for clinicians to systematically report and review the results of diagnostic accuracy trials. This reporting tool consists of a list of 25 items, plus a flow diagram for authors to ensure that all of the pertinent study information is reported accurately (Bossuyt et al., 2003b).

The STARD tool was inspired after the development of the Consolidated Standards for the Reporting of Trials (CONSORT), which was developed in 1998 to guide researchers and authors on how to report the results of their randomized control trials (Smidt et al., 2005). Ochodo et al. (2013) stress that transparency is absolutely required for reporting evaluations of studies that assess diagnostic or medical tests, as they are fraught with inherent biases. As diagnostic accuracy is not comprised of a compilation of predetermined examinations, clinicians need to know which test should be used to accurately provide the best treatment for a given condition. If diagnostic accuracy studies are poorly reported, this inhibits the objective appraisal by the reader and overall generalizability of the study findings. In addition, if only the favourable aspects of the study findings are presented, this may encourage the premature adoption of tests that should not be considered as first-line interventions, resulting in delayed diagnosis and wasting valuable healthcare resources.

1.2.1 The Use of STARD in Radiology

Use of the STARD in the reporting of diagnostic accuracy trials in radiology has been steadily on the rise; however, the adoption of this tool has been less than favourable. Concomitantly, it is important to point out that within the literature, researchers have been criticized for over-interpreting their trial results – the so called “evidence of spin” (Ochodo et al., 2013). Therefore, it is recommended that systematic tools such as the STARD checklist be followed when reporting the study findings so that the results are not misrepresented and clinicians can make treatment decisions with confidence based on the study findings.

Within the field of radiology, certain journals are considered as “high-impact” due to their validity. When I initially began my study, I conducted a brief literature review in September and October 2013 to determine how many diagnostic accuracy trials were reported as per the STARD criteria. The journals included *The British Journal of Radiology*, *European Radiology*, *Investigative Radiology* and *Radiology*. Of the 71 abstracts reviewed, only nine were considered diagnostic accuracy trials whereby the sensitivity and specificity of the index test was compared to the reference standards (Griffiths et al., 2013, Bohte et al., 2014, Fallenberg et al., 2014, Wang et al., 2013, Thieme et al., 2014, Rautiainen et al., 2013, Geffroy et al., 2014, Chen et al., 2013, Cassinotto et al., 2013). Moreover, none of the authors used the STARD criteria, which is alarming as it has been a decade since this reporting tool was published.

This review was repeated in September and October 2015 to determine if the STARD reporting guidelines were adhered to in the same high-impact radiology journals. Of the 326 citations reviewed, 13 radiology articles considered diagnostic accuracy studies were published without the use of the STARD (!!! INVALID CITATION !!!, Cho et al., 2015, Alshamari et al., 2015, Bansal and Young, 2015, Seith et al., 2015, Brucher et al., 2015, Cantisani et al., 2015, Costa et al., 2015, Hauth et al., 2015, Kim et al., 2015, Prummel et al., 2015, Schubert et al., 2015, Yang et al., 2015, Zhang et al., 2015). The search strategies for these searches can be found in Appendix 1 and 2.

Numerous reasons have been postulated about why there's been such as slow adoption of the STARD statement. Of course, it takes time for new guidelines to be adopted by journals and authors. In addition, it has been found that journals have given poor instructions to authors on how to incorporate the use of the STARD checklist (Bossuyt, 2008a). Suffice to say, there is a gap in the literature in that recommendations are necessary to improve adherence to the guidelines when reporting diagnostic accuracy findings. As described by (Ochodo et al., 2013), guidelines such as the STARD are not static, so it is possible that amendments to the guidelines would be beneficial to various subspecialties. Although the CONSORT checklist has been amended twice since its initial publication, the STARD checklist had remained the same until recently, when an amended version of the STARD tool was published in October 2015 (Bossuyt et al., 2015a). Time will tell if the insertion of these additional items, deemed essential, will impact the reporting frequency and rate of adherence of STARD (Leeflang, 2015).

1.3 Research Questions

The purpose of my Doctorate was to answer the following research questions:

a. To determine if the current STARD checklist was satisfactory, specifically for radiology diagnostic accuracy trials?
b. To identify which items of the current STARD require amendment?
c. To identify if any new items needed to be added to the revised STARD are specific to radiology?
d. To determine if the revised tool improves the clinicians confidence level in interpreting radiology diagnostic accuracy trials?

1.4 Research Mindedness

Given the results of my Master's dissertation and my ontological position within our radiology research program, I was aware that current tools such as the STARD were not being utilized to their full potential by radiologists. This knowledge impacted my decision to develop a modified reporting tool with the goal of increasing the quality of reporting and interpretation of diagnostic accuracy trials specific to radiology. As I work in radiology research, I wanted to study a topic that was pertinent to my professional practice as a senior research professional. Given my insider knowledge, I was also aware of which radiologists would be interested in participating in this project, as the results from this study could impact their current thinking or practice. I was mindful of ethical issues such as my own personal perspectives that could have arisen out of the situatedness within the framework of my project. As I had a vested interest in this project, I made every effort to minimize the level of subjectivity that could have arisen between me and those who participated in the project. Those who collaborated on my project provided a great deal of their time and expertise, such that I anticipated the volume of additional work could have resulted in resentment towards the study, and ethical tensions could have arisen (Herr and Anderson, 2005). Fortunately, this was not the case. That said, I did encounter a few ethical issues with a couple of the radiologists that will be described in a later chapter. As action research develops over time, it is difficult to anticipate or mitigate every ethical issue arising. Nonetheless, I was fully committed to comprehensively managing each scenario when it did arise (Herr and Anderson, 2005).

1.5 Action Research

As the goal of my work-based learning project was to develop a tool to improve future practice with respect to the reporting and interpretation of radiology diagnostic accuracy studies, I chose participatory action research as the methodology for my project. This involves research cycling, whereby the validity of an inquiry is achieved via the convergence and divergence of the collaborators as they seek to articulate a “subjectivity-objective reality” (Heron and Reason, 1997:280). My work-based project begun with a needs assessment followed by a Delphi technique, which consisted of several rounds of planning, acting and observing, reflecting and continuing on to the next cycle (Hsu and Sandford, 2007, Koshy et al., 2011). The steps required to conduct my project were completed according to the cycles of action research, whereby the collaborative and reflective inquiry of the working group to learn through action was developed using a participatory paradigm (Koshy et al., 2011).

Concomitantly, the continually collection of data with each cycle of action research could have changed the course of the research and resulted in further ethical situations requiring attention (Koshy et al., 2011). This did not occur. As well, due to the nature of action research cycles, those who agreed to participate could have felt at times that they were “co-constructing the processes of research” (Herr and Anderson, 2005:119). Again, this did not occur. My overall aim of developing a sound research methodology was clearly explained to the working group at the onset of my project, so that they were aware of what they were getting involved with (Costley et al., 2010).

1.6 My Positionality

When I think of my multiple positionalities within my action research project, the most prominent was my "insider/outsider positionality vis-á-vis the setting under study" (Herr and Anderson, 2005:43). This impacted my methodology, which will be described in the methods chapter. I also considered my "hierarchical position or level of informal power within the organization/community" (Herr and Anderson, 2005:44). This ontological awareness also impacted my choice of methodology and how I chose to conduct the action research cycles. Indeed, planning my action research project within my department was accomplished using these multiple positionalities, as well as my sense of research mindedness from years of experience working in research that culminated with the recent completion of my Master's in clinical research. As a novice action researcher, I was learning a lot as I went along, which in some respects was rather destabilizing, as my role within the department had changed when I started conducting my study (McNiff, 2002). Reflecting back on my role as Principal Investigator for this action research study, two of my greatest concerns were ensuring a high physician responder rate and determining the best way to conduct my study within our department. These are described in further detail in my methods chapter.

Within my current position as the Clinical Research Program Manager for the department of radiology I was quite involved with all of our on-going departmental research, which was challenging and exhilarating at times. This awareness was very important to me as it impacted my ontology and epistemology with respect to increasing my knowledge and reflective practice within my research career. Our department of medical imaging is robust, conducting many clinical trials including prospective studies, retrospective chart reviews and/or individual case studies; many of these studies result in publications. In addition, we have conducted numerous diagnostic accuracy trials whereby the outcome of one or more tests (index test) is compared to the outcome of the reference standard (gold standard) to confirm the diagnosis under study in a particular study population (Bossuyt, 2008b). For example, if radiological researchers were comparing computerized tomography (index test) to abdominal ultrasounds (reference standard) to determine the sensitivity and specificity of detecting appendicitis, this would be referred to as a diagnostic accuracy trial.

1.7 Use of STARD within our Department of Radiology

Although our department has participated in many research studies, the STARD statement has not been frequently been used. Nonetheless, I have recommended using the STARD tool whenever possible. For example, in one particular diagnostic accuracy study the Principal Investigator was planning to evaluate the diagnostic performance of volume subtraction computerized tomography angiograms (index test) to digital substraction angiograms (reference standard) in follow-up with patients who had undergone the surgical clipping of cerebral aneurysms. When we received our letter of concern from the research ethics board (REB) with respect to this application, I recommended to the investigator that he respond to the REB with reference to the STARD checklist, as several of these items were relevant to his study.

In another scenario, a retrospective analysis was performed to study the diagnostic accuracy of magnetic resonance imaging for local staging of prostate cancer, whereby pathological results were compared to clinical grading parameters. The objective was to determine the accuracy of prostate tumour detection by considering the grading of magnetic resonance imaging compared to the grading parameters of pathological and clinical detection. Although the study was a diagnostic accuracy study, the title of the study did not include the words “sensitivity and specificity,” nor “diagnostic accuracy.” The first item of the STARD checklist recommends using the Medical Subject Headings (MeSH) of “sensitivity and specificity” in the title of the article, as this will enable the publication to be found through automatic literature keyword searches (Bossuyt et al., 2003a). In this particular paper, the term “correlation” was used rather than “sensitivity and specificity” or “diagnostic accuracy,” so it would have been difficult to retrieve this article when conducting a literature search.

Another diagnostic accuracy study was conducted to determine the diagnostic accuracy of the interface sign (presence or absence of India ink at the margins of the splanchnic vessels) for the diagnosis of vascular invasion in pancreatic adenocarcinoma with a chemical shift MRI. As with the studies already mentioned, certain items from the STARD checklist were adhered to, but the STARD statement itself was not referenced. It is important to point out that the STARD checklist was not intended for retrospective chart reviews; however, given that many of the reviews in radiology compare the diagnostic accuracy of two tests, I would recommend that the inclusion of many items from the STARD checklist would help decrease the use of selective reporting. I will elaborate more on this concept in the literature review chapter.

It is also prudent to mention that the authors of the STARD checklist do not prescribe how a diagnostic accuracy study should be performed (Bossuyt, 2009b). I would argue that researchers, regardless of their clinical expertise, need guidelines when they are conducting research. This includes prospective and retrospective analyses. A key requirement of radiology training is to develop the skills required to perform and interpret radiological examinations and core radiological literature. Given the dynamic field of radiological medicine, radiologists must be critical thinkers as they analyze the literature. The critical analysis of the literature, which is a fundamental principle of clinical practice, is best achieved when research is deemed to be methodologically sound (Budovec and Kahn Jr, 2010). However, many previous radiological studies have been criticized for poor methodology, including small sample sizes and results that are biased (Wilczynski, 2008). Hence, following a guideline such as the STARD will serve to improve the reporting of diagnostic accuracy studies (Medina and Blackmore, 2007). Promoting the use of tools such as the STARD at the inception of a new study may help minimize the criticism of the results once they are published (Leeflang, 2015).

The STARD checklist is generic in nature, as certain items are not pertinent to the reporting of diagnostic accuracy studies specific to radiology; this was the basis for developing this revised reporting guideline. The development of such a tool would aid radiologists and their trainees when reporting, interpreting, and developing research protocols that are specific to radiology diagnostic accuracy studies.

1.8 Reporting Guidelines

One of the most salient points regarding the dissemination and interpretation of the published literature is the overall methodological quality of the reported findings (Moher et al., 2014c). As the results of diagnostic accuracy studies are often used to guide future patient care, it is paramount that the quality of reporting for the trial results is complete and accurate (Bossuyt et al., 2003b). STARD is an initiative that focuses on improving the accuracy and completeness for the reporting of diagnostic accuracy studies, providing the means for readers to detect any potential bias or lack of internal validity, as well as evaluate trial results for overall generalizability or external validity (Bossuyt et al., 2003b). Given that the current STARD tool is not readily used by the radiologists to its full potential is my rationale for developing a modified tool that may be more beneficial to the field of radiology. Upon future publication of the revised tool, it is hoped that radiologists and their trainees will see its value. If the tool is adhered to when reporting and/or interpreting diagnostic accuracy research, it has the potential to benefit future practice in the field.

1.9 Summary of Chapter 1 and Overview of Remaining Chapters

By way of introduction, in this initial chapter I have described how my professional nursing career began and how transitioning into various professional roles that involved working closely with the neurology medical team impacted my epistemological and ontological stance. Upon reflection, this cognitive awareness inspired me to enhance my professional practice through both formal and informal learning. After working for a few years in radiology research, I began my post-graduate degree, whereby the results from my Master’s thesis led to my doctoral project. The following is a brief overview of the remaining chapters of my dissertation.

Chapter 2 begins with a historical description of a landmark study from 1946 that examined how poorly medical literature was published. The results of this study were significant, and it provided the impetus for recommending statistical input in effort to reduce errors in the reporting of health research. This was followed by the recommendation for peer review prior to publication in order to minimize the incidence of bias. My literature review covers the use of STARD in general, and for reporting on diagnostic accuracy trials specific to radiology since its development in 2003. This review resulted in a list of themes that began to emerge, such as which items of the STARD were not often addressed. The overall lack of adherence to reporting tools in general is a main concern, so recommendations for enhancing their use are provided.

Chapter 3 is dedicated to the methodology employed in my study. This chapter begins by highlighting the significance of how the reflection process impacts a researcher's epistemological and ontological stance when deciding which methodology to use to answer their research questions. The methods used to answer my research questions were determined via action research, whereby eight radiological experts participated in a needs assessment and a Delphi technique in an effort to create the newly revised reporting tool named the RadSTARD (Radiology Diagnostic Accuracy Reporting Tool). Once the tool was created, it was validated with the help of radiology residents and Fellows. Each phase of development and validation of this tool involved critical reflection by the participants, as the radiological experts decided which items should be maintained in the revised tool. The trainees also decided if the RadSTARD aided them in interpreting a diagnostic accuracy article. Mixed methods were used, resulting in triangulation of data. Participatory action research was also chosen to conduct this study as this methodology is concerned with a scientific inquiry that is collaborative, as participants have a vested interest in the project when it is deemed context-specific and applicable to practice.

Chapter 4 describes the project activity with respect to how this study was conducted within the radiology department at The Ottawa Hospital. This study was conducted in two phases. Phase 1 began by determining if there was a need to develop a revised tool, followed by the creation of an amended STARD tool called the RadSTARD. In Phase 2, the newly revised tool was validated. Critical friends and consultants were instrumental in the conduct of my study. Some of the greatest challenges involved satisfying the rigorous demands from my local regulatory ethics board, as well as establishing with the original author of the STARD that I was free to create this revised version. Although the department as a whole embraced this project and were eager to participate, there were some who presented a challenge. These individual scenarios are accounted for.

Chapter 5 provides a description of the project findings for the needs assessment and the results obtained from the Delphi technique. In addition, the results from the validation phase of the revised tool are also included. Mann Whitney and chi-square analyses were performed to determine if there were any significant differences between the radiology residents and Fellows in interpreting radiology diagnostic accuracy articles with the RadSTARD tool. Qualitative analysis was done to determine the usefulness of the RadSTARD tool between the two physician groups. An analysis of these results is provided. Additional data include field notes, comments received from individuals participating in the study, and notes from my reflection journal.

Chapter 6 involves recommendations for the future use of this tool, as well as suggestions for implementing and strengthening the adherence to this tool. Finally, Chapter 7 is an overall summary of my reflective journey and key epistemological advancements I have made as a senior research professional conducting an action research project within my workplace.

2. Chapter 2 – Literature Review

2.1 Introduction to the Reporting of Health Research

Each month, over 70,000 medical articles are deposited into the United States National Library and indexed into PubMed, offering portal access to all health related publications. For decades it's been recognized that the overall quality of health research reporting has remained inadequate (Moher et al., 2010b). In order for clinicians to provide exemplary care to their patients based on medical interventions and treatments published in literature, it is essential that published research adequately reflect how the research was conducted, and conclusions made (Rennie, 2014). Health research that is poorly reported lacks clarity and results in uncertainty when not enough details are provided about how the research is conducted. Clinicians and researchers need to strive for transparency when reporting the results of their research, minimizing bias and enhancing overall integrity (Moher et al., 2010b, Rennie, 2014).

2.1.1 Under the Lens of Health Research Reporting

Before we look at methods for improving the reporting of health research, it is important to provide a background on how the mounting evidence of poorly reported publications began. How did it come to be that clinicians and researchers could publish the results of their research in a manner deemed inadequate? In 1946, *The Journal of the American Medical Association* (JAMA) published a landmark paper by Dr. Stanley Schor and Irving Karten, a medical student from Chicago, on their analysis of a random sample of medical reports published in 10 of the most prominent medical journals. Based on their review of the so-called “analytical studies,” 12 types of statistical mistakes were identified, rendering 73% of the reports invalid. In addition, of the 10 journals reviewed, none of their analytical studies were considered acceptable 40% of the time. Due to the recent introduction of computers, the authors speculated that more scientists were publishing and disseminating poorly written research inundating the medical literature. They were absolutely right (Rennie, 2014:xiv)!

Likewise, two of the journals had no acceptable reports. The article also included the results of an additional review, whereby of the 514 manuscripts reviewed, only 26% were found acceptable. However, they also noted that with the guidance of a statistician, the overall acceptance rate rose to 74%. As a result, they recommended that statisticians be part of the whole study design team and provide data analysis for publication. For journal editors, the paper by Schor and Karten became a signpost amongst a plethora of poorly written medical literature, and ultimately lead to the establishment of a Peer Review Congresses. To ensure that the Congress would be beneficial to authors, it was initially restricted to reviewing research reports (Rennie, 2014).

This scrutiny of the medical literature was being observed by Iain Chalmers and his group in Oxford. Chalmers and his group were working on methods to refine how the health sciences were reported and interpreted, ultimately leading to the creation of the Cochrane Collaboration (Rennie, 2014). The Cochrane Collaboration was officially founded in 1992, and named after the British epidemiologist Dr. Archie Cochrane. Membership is comprised of senior scientists, experts in the field of epidemiology and scientific reporting, and membership has grown to over 1,000 members with collaborating centers all around the world. The original mandate of the collaboration was to establish methods for evaluating evidence-based medicine, including the reporting of randomized, controlled trials. This mandate was built on the premise that the provision of health-care will be enhanced if the published medical literature is up-to-date and accurate, rather than being based on anecdotal evidence full of conjecture (Godlee, 1994). Biases in the reporting of diagnostic accuracy results has also been noted and will be reviewed later on in this chapter (Bachmann et al., 2009, Lijmer et al., 1999).

With respect to assessing the level of bias in individual studies for systematic reviews, (Lundh and Gøtzsche, 2008) reviewed the instructions provided to those participating in 50 Cochrane Review Groups, which examined the methodological quality of each study. Although scales were used to score individual items, they were problematic, as some recommendations made had no bearing on the level of bias. Hence, the Cochrane Handbook for authors was recommended for use when rating studies distinguishing between a low, medium or high level of risk pending on the number of criteria met for the articles included in their systematic reviews.

Subjecting published reports to the systematic scrutiny of the experts from the Cochrane Collaboration led to the identification of what items rendered more bias in the final report. It was postulated that there might have been a number of reasons why biases occurred, including conflict of interest, or not publishing at all. This prompted changes to the law, now requiring that trials be registered. By publishing research in clinical trial databases, trials would not only be accessible within the public domain; researchers would also be required to post each stage of trial activity, plus final results. Although researchers have been conducting randomized controlled trials for over four decades, the reporting of those results still lapse in certain circumstances (Rennie, 2014). Such lapses in reporting led to the development of more formal recommendations by experts, to be used when reporting the results of randomized controlled trials. In my opinion, this important milestone was truly the inception for the development of reporting guidelines for healthcare research. I will now review the history of key reporting guidelines, provide elaboration on why we need them, and more importantly, identify how they can be improved.

2.1.2 First Reporting Guideline

Although my dissertation primarily focuses on the reporting guidelines for diagnostic accuracy trials, I will first describe how the guidelines were developed. As mentioned, the lack of quality reporting for randomized controlled trials is considered the stepping-stone to the development of reporting guidelines for healthcare research. In the early 1990s, recommendations were made to JAMA by Senior Methodologist David Moher, to publish trial results according to SORT (Strength of Recommendations Taxonomy) recommendations. This initial attempt at establishing coherence in the reporting of study findings was unsuccessful as the guidelines were too rigid. Nonetheless, it recognized the importance for journal editors endorsing such guidelines as this eventually lead to the development of CONSORT (Consolidated Standards for the Reporting of Trials). Because journal editors endorsed such standards, it forced conformity. The CONSORT recommendations were eventually accepted, leading to the development of many more reporting guidelines in various clinical arenas. In 2008, EQUATOR (Enhancing the QUAlity and Transparency Of health Research) was founded, which not only holds most of the reporting guidelines, but also useful resources for clinicians and scientists to refer to as they prepare their manuscripts for submission (Rennie, 2014).

In summary, close to seven decades have gone by since the article by Schor and Karten was published, and although there has been a great effort to enhance the quality of health research reporting, if errors continue to be found in statistical design and analysis, it will increase the likelihood of results being over-interpreted, otherwise referred to as evidence of “spin” (Ochodo et al., 2013: 584, Rennie, 2014). The literature review of my study was conducted three times since the beginning of the study, during which time radiological experts developed a revised tool for reporting and interpreting radiology diagnostic accuracy studies (Bossuyt et al., 2015b). In the newly revised tool, statistical representation is one of the items recommended. The relevance of this item will be discussed in greater detail later in the Study Findings Chapter.

2.1.3 Why Transparency in Reporting of Health Research is Required

Research related to human subjects is published often, and it should be reported accurately as it has the potential to impact those interpreting the literature and providing future medical treatment based on those interpretations. Such publications are typically found in medical journals, and the results of clinical research are important to not only investigators, clinicians, and systematic reviewers, but also future patients (Altman and Moher, 2014).

When it comes to reporting research, the main question is ‘what do readers need to know?’ There are many different interpretations to this question; it depends on the nature of the research, and also on the audience interpreting the literature (Altman and Moher, 2014). Irrespective of the type of literature that one is interpreting, there are three important fundamental questions to keep in mind. First and foremost, are the research findings deemed substantial enough to encourage implementing an intervention? Secondly, which research outcomes were examined? Thirdly, can knowledge translation to patient care occur as a result of the intervention that was studied (Rychetnik et al., 2002)?

Clearly, when research is reported accurately there is less chance of misleading others through faulty interpretation (Altman and Moher, 2014). To put it another way as stated by Montori et al. (2004:1093), “science is often not objective”. The above statement about science may seem a bit counterintuitive but the fact remains that many investigators have invested emotional and personal interests in their research, which is sometimes required to elevate their academic standing within their institution and affiliated universities. Of course, there are more serious ramifications to the above statement, such as trials that are funded by pharmaceutical companies that also play a part in data analysis and subsequent publication on behalf of those who participated in the trial (Montori et al., 2004).

Perhaps the following quote supplied by the International Committee of Medical Journal Editors (ICMJE) summarizes it best with the following statement:

“In return for the altruism and trust that make clinical research possible, the research enterprise has an obligation to conduct research ethically and to report it honestly” (Montori et al., 2004:4)*.*

Concomitantly, it is important that research reports include enough information about the methods and results, so the reader can judge the overall validity and relevance of the study’s findings. Failure of medical journals to provide enough detail on research methodology is deemed as not fit for purpose (Montori et al., 2004).

2.1.4 How to Recognize Inadequate Reporting of Research

There are two main aspects of a journal article that need to be reported accurately, the methods section and the provision of study findings must be completely represented without ambiguity or selective reporting (Montori et al., 2004). Not providing this essential information is clearly a waste of valuable research resources and has the potential to be harmful. The following list, as described by Montori et al. (2004), provides examples of some of the essential items that require accurate reporting:

- Omitting crucial aspects of the study methodology including eligibility criteria, exact details regarding the study interventions, statistical methods and outcome measurements
- Errors in statistical calculations and selectively reporting statistical analyses
- Omitting to report all of the assessed outcomes
- Failure to report adverse events or harm to study subjects
- Misrepresenting data analysis
- Withholding certain data that is later provided for meta-analysis
- Selectively presenting results in the abstract that are incongruent with the main body of the scientific article
- Inappropriately citing other authors
- Over-interpretation of the study findings and abstract

The impact of such omissions is further amplified when over-interpreted results are published in study registries. Although evidence of these omissions has been found in randomized controlled trials, the reporting of these elements is fundamentally important to all research (Montori et al., 2004).

2.1.5 Impact of Inadequate Reporting and Not Publishing

When clinicians decide not to publish their findings, or only selectively publish partial results, it renders negative consequences to the body of evidence-based medicine (Montori et al., 2004). Evidence-based medicine requires the ability to perform effective literature searches that include selecting the most relevant articles, and applying rules that systematically evaluate the evidence to determine overall validity. Therefore, effectively interpreting the literature, and presenting it to others in a succinct manner while managing a patient's problem, is congruent with a critical appraisal of the literature (Guyatt et al., 1992). If the study methodology is poorly reported, it can impede the interpretation of the study findings (Altman and Moher, 2014).

In summary, I have described the impact that poorly reported health research has for all involved including clinicians, trainees, and most importantly, patients. By describing these elements, I am laying the foundation for the solutions recommended by global expert methodologists and statisticians creating reporting guidelines. Therefore, the next part of my literature review will describe how these reporting guidelines have improved the quality and reporting of health research, and identify the gap that remains with respect to reporting diagnostic accuracy studies specific to radiology.

2.2 Efforts to Improve the Quality of Reporting Research

In an effort to improve the transparency when reporting health research, reporting guidelines now exist for randomized controlled trials and other clinical contexts. As of 2014, the EQUATOR Network has posted more than 200 such guidelines (Network, 2009). The benefit of these guidelines is that they recommend the minimum information that should be included when reporting any type of health research. The guidelines focus on the scientific requirements of an article, which complement the instructions from journals for those who were intending on publishing. Although some of the guidelines are generic, other guidelines are pertinent to research conducted by various medical specialties. Each of the guidelines has been developed by a multidisciplinary group of experts who also provided the rationale for including each item of a particular guideline (Altman and Moher, 2014).

The EQUATOR Network has also included reporting guidelines for qualitative research, as some publishers (i.e., BioMed Central) and journals (Journal of Advanced Nursing) reflected on the need for enhanced rigor when reporting qualitative research (Barbour, 2001). These guidelines were deemed as providing essential information, similar to that of reporting guidelines for randomized controlled trials (Moher et al., 2010a). Although guidelines for qualitative research have been recommended to improve the overall quality of research reporting, there are some who caution that the checklists are overly prescriptive (see below for more information about checklists). As Barbour (2001) argues, subjecting authors to strict conformity of a checklist to increase the overall rigor of reporting is similar to "the tail wagging the dog" (Barbour, 2001:1115). In other words, such conformity could be a deterrent to following reporting guidelines and even publishing overall. The concept of enhancing rigor in the reporting of all health-related research is an important concept that I will address later on in this chapter.

Irrespective of the debate with respect to using a checklist in reporting qualitative research, many medical journals have endorsed the use of generic reporting guidelines, as they ensure all the essential information is included pertaining to methodology, validity of study findings, and generalizability (Altman and Moher, 2014, Barbour, 2001). Although many medical journals such as the Lancet, JAMA (Journal of American Medicine), NEJM (New England Journal of Medicine), BMJ (British Medical Journal), and Family Practice have encouraged adherence to these guidelines, they are not the panacea expected by the scientific community that created these guidelines (Altman and Moher, 2014, Dunt and McKenzie, 2012). For example, a methodological systematic review was conducted whereby 101 reporting guidelines from 1990 to 2011 were assessed to determine if there was a relationship between the completeness of reporting of health research and a journal's endorsement of the reporting guideline (Stevens et al., 2014). The primary outcome of this analysis was defined as authors completely reporting all of the items within a particular guidance checklist. Secondary outcomes included assessment of the overall methodological quality and any incorrect use of the guideline. The authors concluded that there was insufficient evidence to support any relationship between a journal's endorsements of reporting guidelines enhancing the overall reporting of health research. They recommended future prospective analysis via a controlled study (Stevens et al., 2014). This lack of endorsement of reporting guidelines will be addressed again later on in this chapter.

2.3 Developing Reporting Guidelines

As previously mentioned, the first reporting guideline was the CONSORT, which was developed in follow-up to Moher's suggestion to use the SORT, but this did not happen. The successful adoption of the CONSORT's reporting guideline ignited awareness of the importance on the need for complete and accurate reporting of research within the academic community, which has since inspired the development of more than 200 reporting guidelines. Although these developments are very positive, it was also recognized that there was insufficient guidance with respect to developing reporting guidelines (Moher et al., 2014b).

In the quest to harmonize the development of reporting guidelines, senior methodologists created a checklist that is recommended (Moher et al., 2014b). This checklist is comprised of five essential steps, with various items recommended for each step as the guideline is developed. For example, the first step involves identifying the need for a guideline, which could involve developing a whole new guideline or perhaps extending an existing guideline. At this stage, it's also important to identify any potential sources of bias within the various study reports. The second step, termed the pre-meeting activity, involves conducting a Delphi exercise with face-to-face meetings where meeting items are discussed, and the results of the Delphi exercise, shared. This exercise would be conducted by a session Chair. Further consensus meetings would be held whereby the items to be included in the checklist would be listed, and a flow diagram developed. Post-meeting activities would include the development of an explanatory document in the discussion of publication strategies, whereas, post-publication activities would include developing a website to post the guideline and encourage the guidelines endorsement (Moher et al., 2014b).

Although many of these recommendations were followed for the conduct of my study, certain items of Moher's checklist were not adhered to given my choice of methodology. Rationale for this methodology will be discussed in the Methods Chapter.

2.4 The STARD (Standards for Reporting of Diagnostic Accuracy Studies) Recommendations

STARD is an initiative that focuses on improving the accuracy and completeness for the reporting of diagnostic accuracy studies, providing the means for readers to assess for any potential bias or lack of internal validity, as well as the means to evaluate trial results for overall generalizability or external validity (Bossuyt et al., 2003b). In general terms, the diagnostic accuracy of a test determines its ability to distinguish those with the target condition versus those without. The diagnostic accuracy of a test can be expressed in numerous ways including sensitivity and specificity, positive and negative predictive validity, likelihood ratios, diagnostic odds ratios, or receiver operator characteristics (ROC) curves, providing multiple positive cut-off values are provided (Bossuyt, 2014).

In diagnostic accuracy studies, participants meeting the eligibility criteria are consecutively enrolled into the study and exposed to the index test (to test under study) and the reference standard. This is done to determine the best diagnostic tool in establishing the presence or absence of a target condition. In situations where the reference standard is considered error-free, this is also referred to as the gold standard (Bossuyt, 2014). When conducting diagnostic accuracy studies, the provision of a reference standard or gold standard test that is error-proof does not exist (Bertens et al., 2013). This potential problem of reference standards is of particular interest to diagnostic accuracy for researchers needing to choose an alternate reference standard, which was an item that was studied with the cohort that participated in my study.

2.4.1 STARD Compliance

Based on my review of the literature, one of the key themes identified is physician hesitancy to adopt the use of reporting tools such as the STARD (Wilczynski, 2008, Smidt et al., 2005). A review done by Siddiqui et al. (2005) evaluating the use of STARD in the reporting of ophthalmology diagnostic accuracy studies revealed that only half of the STARD items were used. Arguably, conducting this review only two years after STARD was published was probably too early to confirm compliance of its use. (Smidt et al., 2006) conducted a similar review in journals rated with an impact factor of 4, and published pre (2000) and post (2004) STARD, and found that there was only a slight improvement. The review also compared the quality of reporting of diagnostic accuracy articles published in journals that endorsed STARD, versus those that didn't. The results for the reporting of STARD items were similar to those for "adopting journals and the non-adopting journals", which was noted as concerning (Smidt et al., 2006:796). Although this review was also conducted only two years after STARD was published, the lack of adherence to this reporting recommendation is in stark contrast to the adoption of the CONSORT (Consolidated Standards of Reporting Trials) tool which showed improvement in the reporting of randomized controlled trials two years after it was published (Smidt et al., 2006).

Smidt et al. (2006) postulate that the reason for this is that clinicians are more accustomed to the research design requirements for randomized controlled trials, versus diagnostic accuracy trials. I would agree with this, and recommend that clinicians use tools such as the STARD at the protocol development phase of their diagnostic accuracy trial as well. Utilizing standardized reporting tools when developing research protocols was also recommended by Boursier et al. (2015) when they developed a new tool to evaluate liver fibrosis studies.

Another glaring issue is the instructions that editors from STARD-adopting journals provide to those submitting diagnostic accuracy articles for publication. For example, the Journal of Neurology recommends that the STARD checklist be followed when submitting an article for publication, and if their article is accepted, a flow diagram must be provided. Alternatively, the Lancet journal editors suggest that only the STARD tool be followed (Smidt et al., 2006). Moher et al. (2011) conducted a systematic review of the methods employed to develop reporting guidelines and how the guidelines were implemented. In general, their results revealed that in addition to the lack of evaluation for the guidelines developed, there was little in the literature describing how the guidelines were developed. Providing this information, states Moher et al. (2011), is highly recommended as it enables the end-user to determine the quality of the guideline developed and to assess the guideline’s robustness and applicability.

Korevaar et al. (2014) recently published a review of 112 diagnostic accuracy articles to determine STARD compliance of reporting in 12 journals with high impact factor. They concluded that, in general, there was evidence of improvement for some items when compared to studies pre-STARD, but that overall, the reporting of diagnostic studies was sub-optimal. Korevaar et al. (2014) recommended improvement was required in information on selection and eligibility criteria of subject recruitment, the details for interpretation of data – specifically blinding of the interpreters, and finally, the confidence intervals describing accuracy estimates required improvement (Korevaar et al., 2014).

In October 2015, a revised STARD reporting tool was published (Bossuyt et al., 2015b). Since it was first published in 2003, the growing evidence of sub-optimal improvement in the reporting of diagnostic accuracy studies lead to the development of this amended reporting tool (Korevaar et al., 2014, Bossuyt et al., 2015b). Of interest, when the initial STARD reporting tool was developed in 2003, it was recommended for use when reporting the results of all diagnostic accuracy studies (Bossuyt et al., 2003a). However, with this revision, the STARD steering committee is now realizing that other groups may want to develop additional guidance or applications of the STARD tool specific to their area of expertise (Bossuyt et al., 2015b).

2.4.2 Moving Forward – How to Increase Utilization of Reporting Guidelines

After reading the literature on the use of guidelines such as STARD, and the lag in acceptance and endorsement of the tool, I was rather taken aback. As I reflect on our research program and my insider knowledge on the use of the STARD, I am keenly aware that the STARD is not referenced as readily as it should be given our area of speciality. I am referring to studies that compare the diagnostic accuracy of one radiological imaging test to another, in the diagnosis of a particular condition. On further reflection, on the conduct of my study in comparison to those that had previously developed reporting guidelines (Moher et al., 2011), I am confident that radiologists and their trainees will be interested in the RadSTARD (Radiology Standards for Reporting Diagnostic Accuracy Studies), since the development of this tool was developed by radiologists. These methods are fully described within my dissertation (Chapter 5), plus the tool was validated with radiology residents and Fellows.

However, one of the salient issues is how to ensure RadSTARD will be used in the future. What can I do differently than my predecessors who reported such a lacklustre response to the adhering of the various reporting guidelines that are available today (Moher et al., 2011)? Interestingly, O’Leary and Crawford (2013) suggested that allowing potential authors to submit their articles via a web-based platform that confirmed whether the items of the reporting guideline were met may enhance adherence. When the RadSTARD tool was created, a 2-page summary of the tool was also written (Appendix 3). At the time, this was done for the radiology residents and fellows to refer to in addition to the lengthy elaboration document developed (Appendix 4). The RadSTARD 2-page summary briefly lists which items are recommended when reporting the results of diagnostic accuracy studies specific to radiology. This list alone could serve as an application for download to one’s iPad or iPhone. In this digital era of communication, providing this option to physicians may be of interest given that many institutions supply medical faculty and trainees with iPads. Therefore, having the RadSTARD so readily available may enhance its utilization.

2.5 Objectives - Research Questions

The purpose of my work-based Doctorate was to answer the following research questions:

a. How does the current STARD checklist enable radiologists the ability to assess for potential biases and generalizability from published diagnostic accuracy trials and could this be improved when using an amended version?

b. How could STARD items be amended to improve applicability and specificity to radiology?

c. In what ways are clinicians' confidence levels changed when interpreting radiology diagnostic accuracy trials when using the amended STARD tool?

2.5.1 Review of Knowledge and Information

As described by Costley et al. (2010), conducting network-based projects to develop knowledge to solve a practical concern is in contrast to the conventional dissertation, since the goal of network-based projects is to identify gaps in knowledge. Henceforth, the following is a description of how my literature search was conducted to identify articles relevant to the research questions. In addition, based on the results of my literature review, I review knowledge that pertains to the research questions.

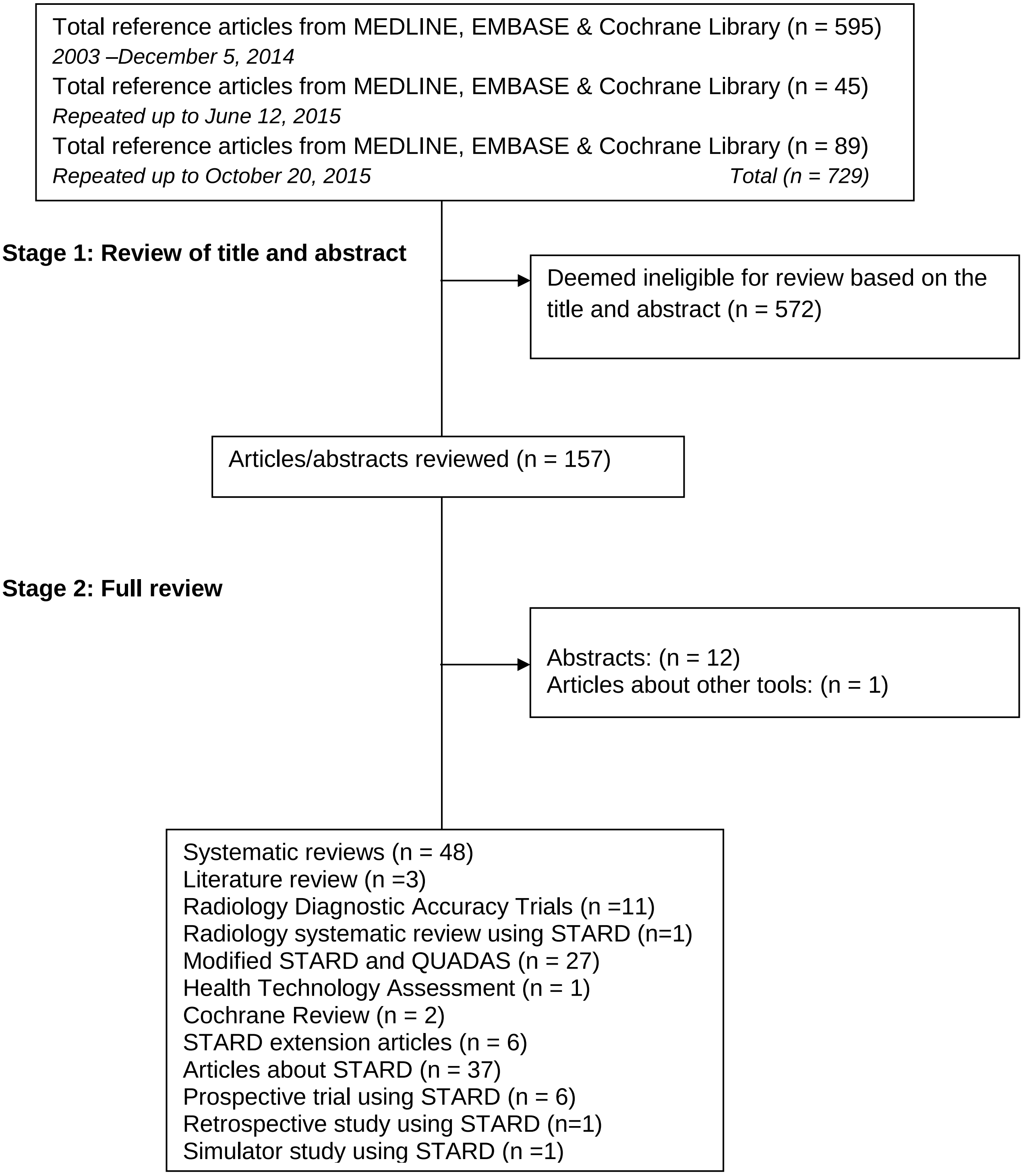
2.6 Literature Search

The literature review of my proposal was conducted by a library scientist at The Ottawa Hospital and consisted of the databases MEDLINE and EMBASE. The Cochrane library was employed for articles relating to the STARD tool and diagnostic accuracy studies published between 2003 and December 5, 2014 (Appendix 5). In MEDLINE and EMBASE, a combination of keywords was used to capture STARD and all its variants and database specific validated headings and keywords for diagnostic accuracy (sensitiv* or diagnos* or di.fs or predict* or specificity or diagnostic accuracy). The search was limited to English language, and research done with human subjects, for a total of 434 results in MEDLINE and 423 results in EMBASE. For the Cochrane library, just keywords for STARDS and all its variants were searched for a total of 96 results. When combined, the results from MEDLINE, EMBASE and Cochrane totalled 953 references and once the duplicates were removed, there was a total of 595.

The literature search was repeated again for dates 2003 to June 12, 2015. This resulted in 54 new references, and once the duplicates were removed, the total was 43 (Appendix 6). The totals were as follows: MEDLINE: 23, EMBASE: 27 and Cochrane: 4. Note, that the totals in the search strategies numbers didn't add up exactly from the previous ones. In December, when the search was run again in Ovid MEDLINE and EMBASE, there were a lot of duplicate references in the database at that time. Ovid reloads its database at the end of each year, and therefore, during the process duplicate references were created. Once the duplicates were removed this changed the totals.

Table 1 illustrates the search totals.

Table 1: Search Strategy Flow Diagram



Total reference articles from MEDLINE, EMBASE & Cochrane Library (n = 595)
2003 –December 5, 2014
Total reference articles from MEDLINE, EMBASE & Cochrane Library (n = 45)
Repeated up to June 12, 2015
Total reference articles from MEDLINE, EMBASE & Cochrane Library (n = 89)
Repeated up to October 20, 2015 *Total (n = 729)*

Stage 1: Review of title and abstract

Deemed ineligible for review based on the title and abstract (n = 572)

Articles/abstracts reviewed (n = 157)

Stage 2: Full review

Abstracts: (n = 12)
Articles about other tools: (n = 1)

Systematic reviews (n = 48)
Literature review (n =3)
Radiology Diagnostic Accuracy Trials (n =11)
Radiology systematic review using STARD (n=1)
Modified STARD and QUADAS (n = 27)
Health Technology Assessment (n = 1)
Cochrane Review (n = 2)
STARD extension articles (n = 6)
Articles about STARD (n = 37)
Prospective trial using STARD (n = 6)
Retrospective study using STARD (n=1)
Simulator study using STARD (n =1)

2.6.1 Validation of Literature Search Terms

All good research begins with a proper review of the existing published literature, which enables a researcher to either continue on where further research has been recommended, or the purpose of the research is to replicate previous findings. An overall summary of the poorly reported research assessing the use of reporting guidelines such as the CONSORT, STARD, and PRISMA (Preferred Reporting Items for Systematic Reviews) reiterated that non-adherence to the guidelines will continue in published literature that renders poor interpretation of the study findings and replication impossible, which as described by Glasziou et al. (2014), are mandatory elements for scientific progress.

With respect to my research, conducting a thorough review of the literature was challenging, as the MESH terms for diagnostic accuracy trials were not easily identified. Wilczynski and Haynes (2007b) conducted a review of electronic databases to determine if there was any improvement in the quality of MESH terms in MEDLINE and EMBASE databases used for diagnostic accuracy trials since the development of STARD in 2003. Journals that endorsed the STARD checklist, versus those that did not, were also compared for two years, pre and post its development. Although no difference was found for MEDLINE, for journals that endorsed the STARD checklist, there was an improvement noted in EMBASE (p = 0.02).

As reiterated by Wilczynski et al. (2013b), the challenges of retrieving evidence-based literature is compounded by the fact that clinical trials deemed strong represent a small percentage of what is available in MEDLINE. Therefore, they developed search filters for a clinical queries database for retrieving clinically relevant research specifically, plus diagnostic accuracy studies. This update was published a decade after their initial recommendation of clinical query filters for MEDLINE (Haynes and Wilczynski, 2004). As such, the library scientist who conducted the search for my doctoral research used this validated search strategy.

While reviewing relevant literature, I also came across an article published in Radiology wherein Astin et al. (2008) recommend developing search strategies to be utilized within the electronic database MEDLINE to retrieve diagnostic performance studies specific to imaging modalities. Based on the recommendations from this article, the search was repeated rendering a further 42 citations, that upon close review, were all found to be irrelevant to the focus of this study. A description of this revised search can be found in Appendix 7.

The literature research was repeated again from 2003 to October 20, 2015. There were an additional 123 references, and once the duplicates were removed, the end result was 89. The totals were as follows: MEDLINE: 53, EMBASE: 65, Cochrane: 5 (Appendix 8).

2.6.2 Results of Literature Review

Based on three literature searches, 729 abstracts were reviewed whereby, 572 were deemed non-relevant. Of the 157 articles and abstracts that were reviewed, only 11 pertained to radiology diagnostic accuracy with one presented in poster format (Saba et al., 2012, Saba et al., 2011, Malcius et al., 2009, Bardou et al., 2013, Crim et al., 2013, Georgantopoulou et al., 2008, Chen et al., 2015b, McComiskey et al., 2012, Laméris et al., 2009, h-Ici et al., 2012, Tseng et al., 2015, Fratz et al., 2009, Miller et al., 2009). Many authors utilized the STARD tool to conduct systematic or meta-analysis reviews (Sharma et al., 2012, Hellemons et al., 2012, Perry et al., 2010, Cid et al., 2010, Wardlaw et al., 2012, Koh et al., 2008, Manchikanti et al., 2009, Miller et al., 2009, Freeman et al., 2009, Mitchell and Coyne, 2007, Zafar et al., 2008, Myburgh et al., 2008, Burch et al., 2007, Martin et al., 2006, Shunmugam and Azuara-Blanco, 2006, Uijl et al., 2005, Siddiqui et al., 2005, Thornton et al., 2013, Haddow et al., 2013, Su et al., 2013, Hing et al., 2009, Coppus et al., 2006, Van Trijffel et al., 2005, Cosse et al., 2014, Chen et al., 2015a, Chiesa et al., 2015b, Chiesa et al., 2015a, Sousa et al., 2015). In addition, one analysis was conducted in conjunction with the CONSORT (Consolidated Standards for the Reporting of Trials) and QUADAS (Quality Assessment of Diagnostic Accuracy) tool (Hall et al., 2008). Other systematic reviews were conducted in other sub-specialities of medicine where it was noted that the STARD tool was modified and QUADAS was also utilized (Wu et al., 2013, Hudon et al., 2011, Thawatchai Leelahanaj, 2010, Mahoney and Ellison, 2007, Chanteau et al., 2006, Cordonnier et al., 2007, Noel-Storr et al., 2013, Raja et al., 2013, Widdifield et al., 2011). Other reviews were done with a modified STARD tool (Lumbreras et al., 2006). The STARD was also used to conduct meta-analyses (Wardlaw et al., 2006, Jahromi et al., 2005, Al-Sulttan et al., Selman et al., 2011, Hewitt et al., 2011). Whereas, other systematic reviews with the QUADAS looked at quality of the studies included for systematic review concluded with the authors recommending that the STARD recommendations be followed for study design and reporting of results (Martin et al., 2006, Selman et al., 2005, Streiner, 2003, Fidalgo et al., 2015, Ligocki et al., 2015, Tsang et al., 2015). There were also systematic reviews done with STARD and QUADAS (Brell et al., 2011, Fontela et al., 2009, Van Trijffel et al., 2005, Bailey and Amre, 2005, Su et al., 2015, Lees et al., 2014, Chavez et al., 2014, Sandrey, 2013, Sousa et al., 2015, Fidalgo et al., 2015, Håkonsen et al., 2015). Rahman et al. (2015) utilized STARD and QUADAS-2 (revised) to evaluate prospective diagnostic accuracy studies that did not include likelihood ratios to measure the impact of false negative sentinel lymph node biopsy in patients with primary breast cancer, versus those who had undergone chemotherapy. STARD and QUADAS-2 (revised) were also used to conduct a systematic review to answer two questions in the paediatric population who received corticosteroids to treat osteoporosis: Could the use of dual-energy X-ray absorptiometry (DXA) determine if DXA could predict risk of fracture in this patient population, and could it be used to determine if bone active treatment of DXA was as responsive when compared to medical intervention and exercise? Wang et al. (2015) concluded that there was not enough evidence to support these claims for DXA.

A review done by Stengel et al. (2005) was a radiological diagnostic accuracy study which referred to both STARD and QUADAS to assess overall methodological design and quality of reporting. Meta-analyses were also conducted with STARD and QUADAS with Shivkumar et al. (2012) recommending using ideal reference standards. One systematic review consisted of adding items to STARD (Mieritz et al., 2012). This study is described further below. A Cochrane systematic review by Flicker et al. (2012) utilized the STARD for Alzheimer's disease, which lead to extensions of the STARD tool and its renaming to STARDdem (Standards for Reporting of Diagnostic Accuracy – Dementia). This revised tool was used along with the QUADAS by (Harrison et al., 2014) to conduct another Cochrane review that assessed the reporting quality of cognitive assessments in this patient population. One review by (Bachmann et al., 2009) used STARD and made recommendations for multivariate adjustments to counter spectrum bias in diagnostic accuracy studies. STARD and QUADAS were also combined with eligibility items from Cochrane to identify the best clinical indicators of ineffective breathing in neonates that lead to a meta-analysis of 6 diagnostic accuracy studies (Sousa et al., 2015). A more recent review of 7 studies, conducted by Sousa et al. (2015), utilized the STARD and QUADAS tool revealing poor reporting of the reference standard.

A review by Goebell et al. (2014) utilized STARD and QUADAS, recommending a tool be developed for diagnostic tumour markers, which supports evidence that one size does not fit all. Similarly, an interesting aquatic study published last year recommended modifying the STARD tool to include study information pertinent to fin-fish pathogens diagnostic accuracy studies to increase the quality of reporting of these studies (Gardner et al., 2014).

A Cochrane review by Pavlov et al. (2015) of hepato-biliary diagnostic accuracy studies was done whereby the authors followed the draft version of the Cochrane Handbook for Systematic Reviews of Diagnostic Accuracy to conduct a systematic review. Several studies could not be included in their review, leading to outcome bias. In order to determine the diagnostic accuracy of transient elastography in staging hepatic fibrosis when compared to the liver biopsy, they recommend that only studies be reported as per the STARD, thereby, reducing the risk of bias.

One particular review done by (Maclean et al., 2014) assessed the quality of reporting of diagnostic accuracy studies with STARD for studies that used cardiac magnetic resonance imaging (CMR), cardiac computed tomography (CCT), and SPECT (single positron emission computed tomography). Overall, they reported an improvement in the quality of reporting of diagnostic accuracy studies in journals considered high impact journals that supported the STARD versus those that did not (p = 0.03). Although these results are encouraging, when the authors compared their analysis to other reviews which reported less successful results, they stipulated that unless the journals render mandatory use of reporting tools such as the STARD, the quality of reporting diagnostic accuracy articles will not improve. This topic of adherence to STARD, and reporting tools in general, will be addressed later on.

The STARD reporting tool was incorporated to report the results of diagnostic accuracy studies of Chinese medicine with human patient simulator (Ferreira and Pacheco, 2015). Interestingly, the authors discussed sample size, which is not an item in the STARD tool. The STARD tool was also used to develop a prospective research protocol (Brat et al., 2015, Geevasinga et al., 2015, Menon et al., 2015, Errico et al., 2012, Wieske et al., 2012, Elikashvili et al., 2014). Retrospective reviews were also conducted with studies that were considered compliant with the STARD (Hiramitsu et al., 2015). Based on this review, themes that emerged from the literature reviews will be described below.

2.7 Literature Review and Use of STARD for Radiology Diagnostic Accuracy Studies

Although my literature search focused on articles pertaining to the development of the STARD initiative and its application to diagnostic accuracy studies, I also read articles pertaining to radiology as a science and how evidence-based medicine has impacted the field of radiology over time. For example, as (Lentle et al., 2007:618) eloquently stated, “As a species, we seem to be fond of reflection at signal times in life” (p. 618). This article was basically a recapitulation of articles published in the New England Journal of Medicine that listed some of the most notable advances in medicine at the dawn of the millennium and within that list was the evolution of imaging sciences. Advancements such as CT (computerized tomography), MRI (magnetic resonance imaging), and PET (positron emission tomography) image guided therapies were highlighted, but the issue that the radiological sciences may have met their limits on the physical attributes used to diagnose tissue was also raised. In other words, the goal of future research should be to take stock of the research done thus far, and determine where the radiological sciences are heading conceptually. It would seem then that diagnostic accuracy and standards to improve the quality of reporting of diagnostic accuracy trials would be a priority for this discipline. The following is a summary of my findings based on my review of the literature.

It is important to highlight some of the salient features of STARD recommended by the scientific steering committee that developed this standard. As described by Bossuyt et al. (2003b), it is recommended that authors refer to the 25-point checklist and flow diagram that includes information on trial methodology such as patient recruitment, the order of intervention for the index test, and reference standard. Of course, there are other salient features, such as how the tests were interpreted and whether a blind was maintained. Failure to meet these standards essentially increases the risk of potential bias to the study and impacts its overall interpretation and the generalizability of the results.

In keeping with the intention to improve the quality of reporting diagnostic accuracy trials, although authors keep referring to the STARD checklist, they fail to report their findings accordingly. For instance, a diagnostic accuracy trial published by McComiskey et al. (2012) compared MRI as the index test to histopathological findings as the reference standard in diagnosing endometrial cancer. Interestingly, the radiologists interpreting the images were not blinded to each others' interpretations of the imaging findings, as both the authors stated that this was unethical outside of a randomized controlled trial. Another glaring omission of this study was that there was no technological detail provided for MRI imaging, paramount for replication of findings. Providing sufficient technological detail is required for both the index test and reference standard, as this will allow others to either replicate the tests or judge the feasibility of conducting the same test within their own practice. If the execution of a test is performed differently, this will vary the diagnostic accuracy results (Bossuyt et al., 2003b).

Conversely, Saba et al. (2011) compared the diagnostic accuracy of MRI and tenderness guided transvaginal ultrasound to diagnose recto-sigmoid endometriosis. The authors did provide thorough details for MRI acquisitions and the study results were illustrated with a flow diagram as recommended by the STARD initiative. The statistics in this article were well done, given the use for sensitivity and specificity, ROC, and kappa. However, there were certain items that were not met on the STARD checklist including a blinded interpreter and providing the necessary years of training for the radiologist that interpreted the MRI findings. I found this very controversial since the same author reported how determining the accuracy of MRI to diagnose endometriosis is reflective of the expertise of the radiologist (Saba et al., 2011). As described by Bossuyt et al. (2003b), the variability of the reader has been noted in previous studies. A certain amount of training is required for the reader to determine if similar results can be obtained in their practice where the level of training and experience is most likely varied.

Malcius et al. (2009) looked at the diagnostic accuracy of different imaging techniques in diagnosing acute hematogenesis osteomyelitis in the pediatric population. Diagnostic accuracy was compared in plain x-ray, ultrasound, bone scintigraphy, computerized tomography, and MRI. In order to establish diagnostic accuracy, the authors looked at sensitivity, specificity, and pause ratio for the various diagnostic-imaging techniques, yet only followed certain aspects of STARD. There was no flow diagram provided, only patient characteristics.

In the same year, Fratz et al. (2009) reported their results of a diagnostic accuracy trial that compared the accuracy of axial slices versus short axis slices with cardiac MRI to determine the best method to measure the right ventricle and left ventricle volumes in patients who had undergone correction for their tetralogy of Fallot. I found this article to be very relevant as it illustrated the importance of including thorough details on imaging techniques of diagnostic accuracy studies specific to radiology. This paper looked at inter-observer variance and concluded that axial slices were more reproducible than short axis slices in cardiac MRI when measuring ventricular volume in patients with tetralogy of Fallot.

Other articles/abstracts reviewed were those identified as diagnostic accuracy trials in other domains of medicine, however, very few features of STARD were reported (Rama et al., 2006, Bardou et al., 2013, Raja et al., 2013). Within the literature, there are reviews done both pre and post development of STARD, in order to determine if there's been an improvement in the quality of reporting diagnostic accuracy studies. Wilczynski (2008) analyzed the literature and found that, overall, the quality remained the same when comparing the two time intervals. One limitation of this review was that the author restricted analysis to items 13-25 of the STARD checklist. In closing, since the development of STARD there is evidence of adherence; however, only certain items have been reported, thus leaving room for improvement.

2.7.1 Systematic Reviews using the STARD

Based on the results from the literature review, as mentioned, the STARD tool has been used repeatedly for systematic reviews. Within my review of diagnostic accuracy trials that had STARD listed in their abstracts, QUADAS (Quality Assessment of Diagnostic Accuracy) was also mentioned if the article pertained to a systematic review. In order to assess whether 3 Tesla (strength of magnet) MRI is more beneficial than 1.5 Tesla MRI, Wardlaw et al. (2012) conducted a systematic review of diagnostic accuracy trials that compared this advanced imaging technology. As the authors used elements of both STARD and QUADAS pre-2003, looking at diagnostic accuracy trials before the systematic tools were developed was probably futile. Interestingly, the authors utilized items from both STARD and QUADAS, including methodology, quality criteria, technical factors, subjects, signal-to-noise ratio, diagnostic accuracy, and errors. The authors concluded that diagnostic imaging technology should be evaluated as rigorously as evidence-based pharmacological interventions, thus improving the quality of care provided to patients, and speaking to the need for new knowledge on improving reporting guidelines.

In one review, the researchers developed a new checklist to enable systematic literature reviews on the quality of reporting that evaluates the reproducibility of 3D motion measurement of the lumbar region. The researchers evaluated the literature with both STARD and QUADAS to develop a new checklist comprised of 4 domains (study population, the test setting, the equipment utilized, and presentation of data analysis) with descriptors added for each domain. Of the 15 articles analyzed, authors found reporting to be incomplete for at least one of the domains. Due to errors in reporting, it was not feasible to interpret reliability and estimates of measurement error (Mieritz et al., 2012).

Flicker et al. (2012) used the STARD checklist in a systematic review of biomarker studies to diagnose Alzheimer's disease and found that there was frequent inconsistent reporting. In particular, items that were poorly reported included lack of blind for the results obtained for the biomarkers and reference standard, poor handling of missing data, and lack of sample size calculations and test reproducibility. This lead to the development of STARDdem, which involved adding extensions of dementia specific rubrics based on expert opinion to the current STARD checklist.

The methods utilized by (Flicker et al., 2012) were somewhat similar to my own methodology, described in the following chapter. Briefly, studying STARD with 10 field experts of dementia is similar to my methodology, except that ours was conducted via the Delphi, which is a format recommended by international methodologists in the field (Moher et al., 2014b). After their tool was iterated several times, it was sent out via their website to an international group of dementia experts for comments. This was most likely done as a way to validate their newly revised tool. Validation of my revised tool was done via a different platform and is described in the Methods chapter.

In a similar fashion, specialists who evaluated diagnostic accuracy studies pertaining to liver disease added extensions to the STARD tool. The impetus for creating a tool specific to liver fibrosis diagnostic accuracy studies was done since the current STARD tool did not take into account imperfect reference standards, spectrum biases, and statistical evaluation for pathological indices specific to liver tests. Eight French experts from the field of liver biology, pathology and biostatistics reviewed ten diagnostic accuracy articles comparing liver fibrosis tests with the STARD checklist. The expert panel was divided into four groups evaluating assigned items from the STARD list with the aim to update or amend the current tool. The experts found that half of the STARD items were partially relevant when evaluating diagnostic accuracy studies specific to liver fibrosis. They recommended amending the current STARD by adding two new additional items, "#12: State if the study is conducted on an intention-to-diagnose basis or if the analysis is per-protocol (i.e. with exclusion of failed/unreliable fibrosis test(s)/reference measurements", and #26: Estimates of cost-benefit" (Boursier et al., 2015:809). In addition, 42 new sub-items were added to the new tool. An accompanying glossary with examples for each item specific to liver fibrosis was also created and was called the "Liver-Fibro-STARD standards" (Boursier et al., 2015:809).

Once the new tool was created, the same panel that created the tool evaluated ten new articles. The purpose was to compare the Liver-Fibro-STARD to the STARD tool, rating each item of the STARD tool as either reported or not, and each item of the new tool as reported, not reported, or non-applicable. Seven international experts rated the same ten articles with the goal of establishing inter-reproducibility for the new tool, resulting in discordance between the internal and external evaluators (Boursier et al., 2015).

I would argue that asking the same experts who created the Liver-FibroSTARD standards to also evaluate the tool could have resulted in bias in their interpretations for the new tool. I agreed with their methods of validating the new tool with the external panel, which lead to further refinement of the tool (Boursier et al., 2015). Oddly enough, validating reporting guidelines seldom occur, as such evaluations are difficult to execute (Moher et al., 2014c). My study did include a validation phase, although it was performed by radiology residents studying to becoming radiologists, and Fellows whom are radiologists sub-specializing in a field of their choice. Hence, I did not ask the expert panel who collaborated in the creation of RadSTARD (Radiology Standards for Reporting of Diagnostic Accuracy) to evaluate it later. I also did not ask the radiology residents and Fellows to compare RadSTARD to STARD as I was measuring their level of confidence when interpreting the diagnostic article with the revised tool. Rather, they were provided the STARD tool, elaboration document, and an article that was reported according to STARD, as a reference only. Further refinement to RadSTARD from radiologists is expected upon dissemination of the tool.

2.7.2 Themes Identified – Key Issues

Based on the results of my literature review, there were several themes that emerged as essential to accurately reporting the results of diagnostic accuracy studies. When evaluating the accuracy of a diagnostic test, one of the main crucial elements is the test's ability to distinguish those with or without the target condition. Although for many years the sensitivity and specificity of a test were considered stable test properties, evidence has shown that the sensitivity and specificity of a test will vary among subgroups based on the results of previous test results, previous co-morbidities and patient gender. Therefore, this source of bias has impacted the results of diagnostic accuracy studies exponentially in the last few decades (Bossuyt, 2014).

The following table is a summary of some of most salient items that were not reported in radiology diagnostic accuracy studies/abstracts when evaluated with the STARD checklist.

Table 2. Radiology Diagnostic Accuracy Studies included in this review

Authors and year	Journal	Design	Location	Participants	STARD items not adhered to	Reference Standard
Lameris et al., 2009	BMJ	Prospective Cohort	Emergency Department	1,021 patients with non-traumatic abdominal pain	• No rationale for cut-offs for index test and reference standard • Selection bias – study population chosen	Reported
O-H-Ici et al., 2012	Journal of Cardiovascular Magnetic Resonance	Prospective Cohort	General Infirmary	46 patients with acute myocardial infarction	• Flow diagram with patient details • Severity of target condition not provided • Cross-tabulation of missing results • Estimates of variability amongst subgroups • Estimates of diagnostic accuracy	Reported
McComiskey et al., 2012	International Journal of Gynaecological Cancer	Prospective Cohort	Cancer Units	213 women with endometrial cancer	• Technical specifications not provided for index test • Cut-off values for index test not provided • Radiologists not blinded on purpose • Years of training for the radiologists • Time interval between index test and reference standard	Reported
Fratz et al., 2009	American Journal of Cardiology	Prospective Cohort	Cardiac imaging referral center	46 patients with corrected tetralogy of Fallot	• Confidence intervals – measures of diagnostic accuracy not provided. • No flow diagram • No description of those with or without the target condition • Adverse events • Estimates of statistical uncertainty, test reproducibility • How missing data was handled	Reported

Authors and year	Journal	Design	Location	Participants	STARD items not adhered to	Reference Standard
Crim et al., 2013	Skeletal Radiology	Retrospective Cohort	Not indicated (Podium presentation)	55 patients with chondrosarcomas	• Technical specifications of index test • Whether the interpreters were blinded • Years of training • No confidence intervals	Reported.
Malcius et al., 2009	Medicina	Prospective Cohort	Pediatric surgery ward	183 pediatric patients with suspected acute hematogenous osteomyelitis	• Technical specifications • Reference standard rationale • Training and how tests were interpreted – level of blind • Test reproducibility • Estimates of variability for diagnostic accuracy	Not reported.
Saba et al., 2011	Journal of Magnetic Resonance Imaging	Prospective Cohort	Gynecological unit	59 patients with suspected deep pelvic endometriosis	• Adverse events • Indeterminate results.	Reported
Saba et al., 2011	European Journal of Radiology	Retrospective Cohort	Medical center	30 patients with suspected endometriosis	• Rationale for reference standard. • Distribution of disease severity • Cross tabulation of results • Estimates of variability between subgroups	Reported
Georgantop-oulou et al., 2008	Gynecological Surgery	Prospective	Menorrhagic Clinic	140 patients women with menorrhagia	• Training of the operators was not specified • Cross tabulation of results • Test reproducibility • Estimates of variability of diagnostic accuracy between subgroups of participants	Reported study as per STARD tool.

Some essential information was omitted pertaining to how diagnostic accuracy studies were conducted (technical specifications), level of blind when images were interpreted, lack of information pertaining to estimates of diagnostic accuracy (confidence intervals) – some of the most salient STARD items that were not reported in this review (Crim et al., 2013, Saba et al., 2011, Malcius et al., 2009). One article was reported as per STARD recommendations; however, not all items were included in the article (Georgantopoulou et al., 2008). I found that not providing details on the level of training for the operators rather concerning as, without this information, it is difficult to ascertain replication in one's own institution. As the clinical practice of radiology evolves when incorporating research-based evidence, it is imperative that published research such as diagnostic accuracy studies are reported accurately (Brealey and Scally, 2008).

Within the realm of evidence-based radiology, radiologists are required to interpret imaging with a level of comprehension that includes incorporating evidence as provided in the literature. Therefore, aspects such as the description of technical expertise are crucial to evidence based radiology (Sardanelli et al., 2010). The methods employed in a study strengthen its validity. As it is impossible to eliminate all biases and confounding variables, all measures to reduce their incidence of occurrences should be constrained when possible (Budovec and Kahn, 2010).

As per Bachmann et al. (2009), the two main sources of bias that can occur in diagnostic accuracy research are spectrum bias and test review bias. When test parameters are skewed due to variance in the study populations this is referred to as spectrum bias. Whereas when there is a lack of information provided to those interpreting the test results, this will render test review bias. Bachmann emphasizes the importance of adhering to both STARD and the QUADAS when reporting the results of diagnostic accuracy studies. However, for the purposes of this study, only the STARD will be focused on. The authors advised adding an additional statement to item 23 of the STARD tool: “Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done” (Bossuyt et al., 2003b:16). To address the issue of bias, Bachmann et al. (2009) recommend that authors provide additional information on any subgroup assessments by providing multivariable statistical adjustments. Likewise, previous authors have analyzed the quality of reporting of diagnostic accuracy literature with a modified STARD tool that did not include this item (Walther et al., 2014, Wilczynski, 2008). Alternatively, Gatsonis (2003) purports that most items of STARD pertaining to diagnostic imaging should include the requirement to report the degree of inter-observer variability between readers. An assessment of this feature is very pertinent to diagnostic imaging, encouraging that any subgroup analysis also be reported by the cohort. Further discussion on this item will be reviewed again in the Chapter on Study Findings.

2.8 Enhancing rigour in the reporting of all health related research

In summary, this literature review included reviews in radiology and the results of diagnostic accuracy studies completed in other domains of medicine. Determining the accuracy of a diagnostic imaging test requires studying the test in a cohort of patients with the suspected medical condition under study, as opposed to conducting randomized controlled trials. This so-called practice has been termed evidence-based radiology, as it is specific to radiologists. As radiologists interpret imaging, a full comprehension of the implications of their findings in the context of the evidence reported in the literature is required (Sardanelli et al., 2010).

The use of STARD in the reporting of diagnostic accuracy trials in radiology has been steadily on the rise; however, the adoption of this tool has been variable. Concomitantly, it is important to point out that within the literature, researchers have been criticized for over-interpreting their trial results – the so called "evidence of spin" (Ochodo et al., 2013:584). Therefore, it is recommended that systematic tools such as the STARD checklist be followed when reporting study findings so that results are not misrepresented, and that clinicians can make treatment decisions with confidence based on the study findings. The limitations of the current STARD tool for reporting radiology diagnostic accuracy studies was evaluated by those who were involved in amending the tool. In the validation phase, the new tool was reviewed in comparison to the previous STARD by the residents and Fellows in our radiology department.

The next chapter describes how the main research paradigms impacted the choice of methodology to conduct an action research project. In particular, I chose participatory action research, commonly used by healthcare professionals with the goal of improving one's practice. In this scenario, it was my goal to improve the reporting of diagnostic accuracy research specific to radiology. This chapter describes how this goal was met via the reflective cycles of action research.

3. Chapter 3 – Methods

3.1 Introduction

The purpose of this Doctorate was to develop a reporting tool that would increase the quality of reporting of diagnostic accuracy trials specific to radiology. Once the project was approved by Middlesex University and my local institutional ethics board the project commenced. An action research approach was undertaken within the radiology department at The Ottawa Hospital, Ontario, Canada. Where the goal is to improve practice, action research is a particularly well-suited methodology that is frequently utilized within the health sciences (Koshy et al., 2011). This typically entails utilizing qualitative data collection methods such as conducting interviews, maintaining a reflective journal and observation (Meyer, 2000). Similar methods were employed, and as we studied the current STARD (Standards for Reporting of Diagnostic Accuracy) systematic tool, questionnaires were used to obtain feedback required throughout the cyclical stages of the action research project (Bossuyt et al., 2003b).

As described by Grix (2010), it is very important that one clearly comprehends how their perspective can impact the research process. This entails clearly delineating the relationship between one's ontological and epistemological positioning, as the two are separate entities that should not be lumped together. By this, Grix (2010) is referring to the reflection process a researcher undergoes when considering what they want to study, also known as one's ontological positioning. This reflection links with one's epistemology or current knowledge about a particular area of study which influences what type of methodology is needed to generate the knowledge necessary to answer the research question. As an insider researcher, I was aware of how infrequently STARD was utilized in research studies. In deciding how to improve research practice with respect to the utilization of reporting tools such as STARD, I took a “critical stance toward mainstream research” (Morrison and Lilford, 2001:436). By this, I am referring to the fact that action researchers need to consider how to pioneer action research so that it is appropriate for the level of enquiry (Morrison and Lilford, 2001). Therefore, I adopted key theorists such as Jack Whitehead's “I-approach” to action research as my enquiry in an effort to improve practice and develop the action research project (McNiff and Whitehead, 2002:54).

This knowledge impacted my choice of research questions:

1. How does the current STARD checklist enable radiologists the ability to assess for potential biases and generalizability from published diagnostic accuracy trials and could this be improved when using an amended version?
2. How could the STARD items be amended to improve applicability and specificity to radiology?

3. In what ways are clinician's confidence levels changed when interpreting radiology diagnostic accuracy trials when using the amended STARD tool?

Although there are many cognitive and influential models to adopt when constructing knowledge as described by McNiff and Whitehead (2002), there is also a tension within the society of action researchers with respect to concepts employed in the development of knowledge. On the one hand, knowledge that is a "given" is acquired (McNiff and Whitehead, 2002:39), whereas knowledge generated by researchers is accomplished by working through their issues. In my study I began my research with knowledge that was a "given" or acquired with respect to use of STARD within our department (McNiff and Whitehead, 2002:39). Secondly, knowledge was generated from my methods, which included a needs assessment, Delphi technique and validation, elaborated later on in Section 3.6 of this chapter.

Given my current level of research experience, employing action research as the methodology of choice was a natural one since our working group (radiological experts) were accustomed to conducting research in a participatory fashion (Koshy et al. 2011). In this chapter I will describe how traditional research paradigms intertwine with action research throughout the different phases of cyclical enquiry. Details of the actors involved, and their role in the various aspects of this research, will be outlined. Methods of data collection including my rationale for their choice will be reviewed and limitations described. Data analysis and ethical considerations including issues of trustworthiness will be addressed and conclusions provided at the end.

3.2 Research Paradigms

Research paradigms are the foundation of current beliefs and assumptions guiding the conduct of research and interpretation of data. Such beliefs are shared amongst disciplines where scientific inquiry is regulated. Overall, paradigms are defined by their ontological and epistemological stance, as this impacts one's chosen research methodology, resulting in the construction of disciplinary knowledge (Bunniss and Kelly, 2010). Action research is often chosen within healthcare practice where the goal is to improve practice and create change (Koshy et al., 2011).

Historically, thinking, rather than knowing through doing, has produced scientific knowledge. The Cartesian roots of traditional science is based on doubt, "dubito, cogitio, ergo sum", which means the very awareness of knowing and thinking is possible due to one's existence (Reason and Bradbury, 2005: xxv). Such beliefs parallel my earlier statements on the development of knowledge in the action research community (McNiff and Whitehead, 2002). Choosing action research as my methodology was influenced by my ontological and critical stance on the research community, in this case, the radiologist's use of

STARD when reporting and/or interpreting radiology diagnostic accuracy studies. In unpacking this truth, my action research study was conducted in keeping with the critical theory research paradigm.

Within literature and depending on the source, research paradigms are given different terms which tends to confuse this important concept. Main research paradigms include positivism, post-positivism, critical theory and constructivism (interpretivism). Recently, a fifth paradigm was added - participatory or cooperative paradigm (Guba and Lincoln, 1994, Heron and Reason, 1997). This paradigm is congruent with participatory action research and the critical paradigm.

3.2.1 Research Paradigms

The popularity of paradigms has been attributed to Thomas Kuhn's book, "The Structure of Scientific Revolutions" which describes how one's set of beliefs guide the development of knowledge by researchers. Although this landmark book was greatly acknowledged as a cornerstone for guiding researchers, Kuhn was also met with criticism for referencing the term 'paradigm' in more than 20 different contexts throughout his book. Denounced by Morgan (2007:50), he advocates that it may have been beneficial if Kuhn described the degree of commitment and level of agreement required by the researchers within a disciplinary matrix. As this did not occur, scientists talk about paradigms which mean different things to different people (Morgan, 2007). Based on my exposure as a novice action researcher, I concur with Morgan. Although this type of methodology is beneficial due to its flexibility and critical reflection, successfully conducting an action research project required a great deal of commitment from those who agreed to participate in the study.

Within social and applied sciences, some researchers have argued that the definition of research is influenced by one's theoretical framework whereby theories are established to connect relationships between constructs. Sometimes this theoretical framework is referred to as paradigm, which impacts how knowledge is studied and interpreted. Therefore, the first step in conducting research is to choose a corresponding paradigm that describes the intentions and expectations for the proposed research. If a paradigm is not chosen, a foundation is lacking for the choice of methods employed in the methodology section, the literature search, and design of the research project. The fact that paradigms are not always discussed in research, and when they are their definition is not always clear, only adds confusion to this important first step in the research process (Mackenzie and Knipe, 2006).

It's important to note that rarely are ontological and epistemological perspectives discussed within medical literature when research methodology and paradigms are described. This is not always the case however, especially in nursing research where the journals tend to locate themselves within the premise of non-positivistic research and premises are made known to the reader (Bunniss and Kelly, 2010). One exception to this is a publication by McNamee et al. (2009) describing the perceptions of medical students' experience when they were taught the medical legal implications of autopsies. Within this article, the methods section clearly describes how a phenomenological approach was chosen instead of a classical positivist one (McNamee et al., 2009). Otherwise, most medical journals tend to focus mainly on the methods used to collect and analyze data (Bunniss and Kelly, 2010).

Only providing the methods employed to obtain and analyze data is problematic because the researcher's knowledge and epistemological views are not known. For instance, the reader may question if the study was conducted within an ethnographic or critical theory approach? Therefore, it's important for the researcher to define their knowledge and epistemological stance as this impacts the development of the research question (Bunniss and Kelly, 2010). I agree with the recommendation made by (Bunniss and Kelly, 2010), and reflected how the basic premises of the main research paradigms shaped my research questions and choice of research methodology throughout the development and progress of this doctoral study. I also concur that paradigms, or epistemological and ontological stances are not often discussed in the methods section of medical research. In my experience they are not within the research protocol or published results. Alternatively, the methods section is described with precision and governed by rules one would expect to find within medical research. This lack of additional information on one's paradigm and experiences is somewhat of a natural phenomenon. Medical researchers tend to focus on calculating sample sizes and objective measures rather than considering their epistemological stance and the paradigms driving their research methods. I contend that this practice should not continue and when I complete my doctorate, I will add this recommendation to my results upon dissemination as I truly believe this requirement is an integral part of the practice turn (Schatzki et al., 2000).

3.2.2 Positivism Paradigm

In general, philosophies of medical research have traditionally been dominated by positivistic paradigms (Bunniss and Kelly, 2010) while variations of the interpretivist paradigm have also been quite influential. From the 1930s to the 1960s, the dominant epistemological paradigm within the social sciences was positivism, which held the basic premise that the properties of the social world were external to the researcher but could be measured directly by observation. In general, positivism contends the following:

Reality is available via the senses based on what is being heard, smelt or touched. Inquiry is empirical as it is accomplished via scientific observation versus philosophical speculation. The methodological principles of the natural and human sciences share a commonality that deals with facts versus values (Gray, 2013).

The positivist paradigm is therefore referred to as the scientific method as it is based on the assumption that knowledge gained is from observable data resulting in objective reality. The methods utilized in this paradigm consist of quantitative measures whereby consistency between the variables are measured and identified (Koshy et al., 2011, Fox et al., 2007). As described by Solomon (2011:454), evidence-based medicine has been referred to as the "Kuhnian paradigm". In other words, what is admissible as evidence in medicine? What is the hierarchy of evidence? For instance, randomized controlled trials are considered a higher level of evidence than systematic reviews (Solomon, 2011). Knowledge is developed based on ideas of scientific inquiry that is tested and adheres to strict rules. Consequently, some practices of positivism have been met criticism as not all science is based on mathematical formulations. There are still grey areas in science interpreted by reasoning, rather than by direct evidence. Typically, science is built on theory rather than observations, and theory allows for interpretation of observations (Gray, 2013).

By way of example, positivist paradigms can provide the foundation for answering research questions; however, they are rather restrictive given that not all research can be conducted via randomized controlled trials. As a consequence, these methods have been deemed invaluable for alternative research endeavors where the goal is to study "complex, unstable, non-linear social change" (Bunniss and Kelly, 2010:358). So, although approaches to research were originally built under the auspices of positivism, empirical inquiry, experimental designs and inductive generalization, today we tend to operate within a post-positivist paradigm whereby alternative paradigms have emerged (Gray, 2013).

Upon reflection, as our research is conducted within a medical institution, gravitating toward a positivist paradigm seemed natural. However, based on my clinical research experience over the past two decades, I have worked with many alternative research methods and learned that not all science is objective. There are many other methodological alternatives to choose from when conducting research. Due to my experiential knowledge and extensive level of independent study in research sciences, I realized that I entered a post-positivist paradigm. Although some research still falls within the positivism realm, many research initiatives do not. It is this epistemological stance and ontological awareness that guided me towards the choice of action research as the method of conduct my doctoral study. I will now describe the essential elements of post-positivism that served as the paradigmatic foundation required to answer my research questions.

3.2.3 Post-Positivism

The post-positivist position maintains that there is a reality that should be investigated by research. However, given the limitations of the research process, truth unfolds slowly. It also accepts the fact that the researcher's position is not value free since the researcher's background helps to develop both the research and the results measured. Although the aim is to be objective, the reality is that it's not always possible (Fox et al., 2007). The use of alternative paradigms in medical education research are beneficial as they provide reflection for research questions that are somewhat ambiguous and require a certain degree of flexibility in study design. Henceforth, utilizing research methods in medical education research that includes epistemological and ontological premises is appropriate for complex studies that reflect change (Bunniss and Kelly, 2010).

At the end of World War II, post-positivism replaced positivism. Post-positivists maintain that research is influenced by several well-established theories in addition to the paradigm that is being tested. Post-positivist paradigms challenge the original theoretical framework by Thomas Kuhn that knowledge is constructed with the belief that the world is consistent of multiple ambiguous realities where the truths found by one group may not be consistent with those found from another, which aligns with the constructivist paradigm. Post-positivist paradigms explore qualitative methods that result in research findings, and as such, are holistic and intuitive in nature where others have maintained that quantitative methods have been used for data collection and analysis for both positivist and post-positivist research (Mackenzie and Knipe, 2006).

3.4 Action Research and Paradigms

Although Kurt Lewin is thought of as the original pioneer of action research, his theories on progressive educational learning were influenced by philosopher John Dewey (Reason and Bradbury, 2005, Waterman et al., 2000). Lewin advocated that the theory associated with action research experiments must describe how the theory could be linked to the experimental results. Conducting an experiment without this relationship would result in the experiment not making sense. One could still interpret the experiment; however, the interpretation would be more hermeneutic, as Lewin pointed out, because most of the methodological features of the experiment were left out (Gustavsen, 2001). For this to occur, a relationship between each separate variable of the experiment must be linked to theory (Waterman et al., 2000). Lewin believed that knowledge could be created when real-life problems were solved (Herr and Anderson, 2005).

Within action research, there are three post- positivist paradigms: interpretative, critical and participatory. Certain elements of these paradigms justify choosing action research as methodology of choice for this research study. For example, one of the key elements of critical theory is that action research empowers those involved in the study, enabling them to take control of a particular situation and change it for the better. Whereas the interpretative paradigm is more concerned with social ramifications such as the attentiveness required when studying a staff's perception when working with them as an outsider, the participatory paradigm is applied for those aiming to include patient participation in their study (Koshy et al., 2011). Once the revised tool for interpreting radiology diagnostic accuracy studies was developed, the next phase was to validate it and the phases of this study were congruent to the spirals of action research, described later in this chapter. Although I gravitated toward critical and participatory paradigms of action research, I would argue that the underpinnings of the critical paradigm were the predominant methods required to obtain the data I needed to create new knowledge in the reporting and interpreting of diagnostic accuracy research specific to radiology. It's this knowledge and awareness, I am optimistic will lead towards improving the current ontology in radiology research and impact future practice with respect to reporting and interpreting the results of diagnostic accuracy research.

Both quantitative and qualitative methods can be utilized for any of these paradigms even though quantitative methods are traditionally aligned with positivistic paradigms and qualitative paradigms with interpretative paradigms (Koshy et al., 2011). Although action research methodology is more frequently used by social sciences, it is also used in healthcare research (Somekh, 1995, Koshy et al., 2011). Irrespective of the domain of research, the underlying theme of action research is the same: to bring about a change in practice whereby knowledge is generated by integrating research and action (Somekh, 1995). I would like to point out; one of the salient differences in my research is that when conducting action research, practitioners have decided or perceived that change is required. The difference here is that I am the sole researcher that decided that change in the reporting and interpreting of diagnostic accuracy research specific to radiology is required. As previously described, this was based on my prior level of study, experiential and tacit knowledge in radiology research and my overwhelming desire to improve the research process and create a practice turn (Schatzki et al., 2000). In the spirit of research and action research, once this step was clearly defined with the development of my doctoral research proposal, the study progressed into the next phase of the research process. The study was conducted in two distinct phases and within these phases were many cycles synonymous to the "spiral process" of action research (Afify, 2008:156).

3.4.1 Action Science

The term action science is strongly aligned with the work of Chris Argyris whose main concern is the development of knowledge and learning within an organization. Argyris viewed organizations as ”self-correcting systems” whereby communication was integral to organizational change (Herr and Anderson, 2005:14). Argyris gave academic workshops teaching practitioners how to recognize ‘reflection on action’ and encouraging engagement with others to foster the beliefs necessary for successful learning and development within an organization (Smith, 2001). His work incorporated aspects of critical theory such as Habermas' premise that communication established by free and open discussion renders a better argument (Herr and Anderson, 2005).

Argyris also believed that by mainly focusing on solving problems via action research we transition too far away from the scientific dimension of building theories and testing them. According to Argyris, action science consists of “knowledge that is useful, solid, descriptive of the world, and informative of how we might change it” (Herr and Anderson, 2005:14). He argues that the methodology of certain types of action research attempt to conform to the restrictions of research deemed rigorous within the traditional social sciences so much so that the methods become detached from reality and are no longer useful (Herr and Anderson, 2005).

Building on the work of his predecessors Dewey and Lewin, Aryrgis supports interventions that result in change in learning within an organization. These concepts are important to action researchers because they encourage action researchers to solve problems in a way that changes the practice in a non-superficial and temporary way (Herr and Anderson, 2005). With respect to my own study, I concur with Aryrgis' sentiments given that the development of a new tool designed to enhance interpretation of radiology diagnostic accuracy trials, be it in the development of protocols, or recommendations of which items to include when publishing, is being done with the goal of enhancing research practice. The active participation of the residents and Fellows during the validation phase is meant to render a change in the way they report and interpret radiology diagnostic accuracy research in a long-standing, rather than temporary fashion.

3.4.2 Knowledge and Action Research

Although many action researchers have taken the critical approach when conducting their action research, there are few accounts of how reflection and reflexivity occurred (Herr and Anderson, 2005). Jurgens Habermas who purported that the construction of knowledge is never neutral in that there is always a given purpose in mind cited this most often. Habermas believed that communicating within any discipline was distorted since power relations interfere with the act of communication (Herr and Anderson, 2005). Power relations were carefully considered in my methodology for both phases of the action research study. Whether I was inviting radiological experts to participate in the development of an amended version of the STARD tool, or inviting the residents and Fellows to validate the new tool by testing, everything revolved around the consenting process. Given my position within the department, I was cognizant of the power relations between those participating in the study and myself. I knew it would take repeated communications with the working group to get their participation and that I would need to exercise care to encourage, rather discourage, participation. If they chose not to participate, their decision would be respected.

Reflecting on one's professional development has become common practice across a multitude of disciplines. As such, Schön's theory on the relationship between "reflection-in-action" and reflection between "reflection-in-action" is a good segue for introducing the processes that occur when research is conducted within one's practice. Argyris and Schön take reflection one step further when they describe "theories-in-action" as practitioners reflecting on their everyday work and "espoused theory" when practitioners describe the reflection involved with these theories and actions, to others. Such articulation requires reflection and action whereby original reflections are reformulated (Scott et al., 2004:56-57).

Reflection on practice is further defined by "single loop" learning whereby the loops of action research link to the reflection involved within the framework of an organization which Argyris and Schön refer to as "governing variables" (Scott et al., 2004:57). These variables are the boundaries that are agreed upon as acceptable limits. Any action that occurs within this preset determined boundaries may result in compromising the governing variables (Smith, 2001).

According to Argyris and Schön, detecting and correcting errors results in learning. If something were to go wrong, strategies would be developed to work within the governing variables. This would include questioning the underlying goals and values, consistent with "single-loop learning". Alternatively, one could question the underlying governing variables by subjecting them to critical scrutiny, commonly referred to as "double-loop-learning" (Smith, 2001:8) Such scrutiny could render altering the underlying governing variables resulting in organizational learning. Essentially, the existing goals, values and strategies taken for granted within a framework result in single loop learning where the goal is to make techniques more efficient. In this scenario, reflection is aimed at enhancing the efficacy of strategies. Alternatively, double loop learning questions how learning can occur within the framework to enhance the underlying goals, values and strategies (Smith, 2001).

The goal is to develop models that impact double loop learning by either enhancing or inhibiting the process, as described by Argyris and Schön. The authors subscribe to the fact that people tend to employ a theory-in-use when addressing a particular problem. Such learning is called Model I, and inhibits double-loop learning; whereas Model II learning is more concerned with governing variables and their affiliation with the theories-in-use resulting in enhanced double-loop learning (Smith, 2001).

Distinguishing between these models of learning is beneficial as it relates to the study activity associated with a professional doctorate. At the core of these learning models is a reflective activity that occurs either proximally or distally from the object being reflected upon and is described as being either reflection-in-action versus reflection on reflection-in-action. This reflection is nonlinear, depending on the type of reflection occurring in relation to the activity, reflecting becomes more of a complex synthesis. Indeed, the activity associated with this study was complex allowing for meta-reflection. I have found this to be very beneficial since the overall goal of a professional doctorate is to become a reflective practitioner (Scott et al., 2004).

The main goal of action research is to produce practical and useful knowledge to those within a particular discipline. Therefore, through the creation of newer forms of comprehension, action research consists of working towards outcomes that are practical. Action research without comprehension and reflection is visionless (Reason and Bradbury, 2001).

3.4.3 Action Research Types

Before listing the various types of action research, I would like to review some of the basic research tenants of action research. Conducting action research includes systematically collecting data from planned interventions, whereby the data is analyzed and implications from the data findings are reflected upon for additional observation and action. Depending on the results of data analysis, a further cycle of planning, interventions and actions may be initiated. Given that action research processes are subject to change direction based on analysis of a previous cycle allows for flexibility in methodology. Such actions are performed based on specific contingencies within the research environment. Due to the less than predictable nature of action research, when compared to other research methods, action research distinguishes itself by spiral of cycles. The cycles involve “planning, acting, observing and reflecting” within the context of one's research environment (Burns, 2005).

One of the most popular versions of the action research cycle was developed by Kemmis and McTaggart who purported that there were four essential steps required between each iteration of self-reflective spirals conducted according to the goals and purpose of the research study. These steps include planning, action, observation and reflection (Burns, 2005).

Although I chose action research as my methodology, Bradbury and Reason (2006:xii) view action research as “orientation toward inquiry”, an intention to create engagement through the questions posed in practices and evidence gathered. I concur with their concept of action research, adding that throughout the orientation towards inquiry, significant learning and enlightenment occurs as the actors reflect on their responses, leading to change. However, the basic underlying component of action research methodology is the collaboration between practitioners and researchers. Many interpretations of action research exist and irrespective of use, action research requires a union between praxis and theory. Praxis describes the cyclical process of experiential learning and is often used by educators. It has also been described as a form of critical thinking as it comprised of critical reflection and action (Afify, 2008).

There are three types of action research: technical, practical and emancipatory where the aim of the research and the goals of the researcher or facilitator are clearly delineated with the roles of the participants. The common thread between these three types of action research is the idea of collaboration as the key to the research process (Afify, 2008). As a novice research practitioner, I reflected on this underlying theme of action research and set out to conduct the study within my workplace in a manner to foster collaboration. I decided then that my best approach would involve choosing participatory action research methodology.

3.5 Critical Theory

When studies are conducted in the real world, the ‘action’ component involves exploring how an action can be facilitated to create change or influence particular policies and practices. Some action processes encourage emancipatory or an empowerment purpose, which is a dominant theme in feminist research where the goal is to help members of an oppressed society take charge of their lives. Such change can be done directly, as change that occurs due to the study itself, or indirectly through policy amendment (Robson, 2002). The emancipatory nature of action research play a dual role as it can help create new knowledge that is practical, as well as the abilities required to create that knowledge (Reason and Bradbury, 2005). Marcuse and Habermas were critical theorists who followed in the footsteps of Marx, believed that the task of the philosopher was to change the world rather than understand it. Their theories became important in the establishment of authoritarian power within societies. For example, critical theory was associated with welfare projects, political movements, dimensions of social oppression, disabilities, democracy and social justice with the goal of establishing equality (Robson, 2002).

The underlying commonality with these approaches is that the researchers and the participants expect active participation. In order to achieve active participation the following key features must be followed:

- The overall goal has to evoke a practice change
- The researchers have to present an action agenda to the participants
- Power relationships have to be minimized by providing a scenario for participants free from constraints
- Self-development and self-determination have to be encouraged
- Debate and discussion have to be encouraged and fostered to enhance a change in practice
- Active collaboration by the participants in all phases of the research has to be ensured (Robson, 2002)

3.6 Participatory Action Research

Participatory action research (PAR) is a commonly chosen methodology amongst healthcare professionals for conducting research. PAR methods often involve observing participants or requesting respondent’s feedback to questionnaires or interviews. Although this is truly characteristic of PAR methodology, these methods have been criticized for not being “action orientated” enough to enhance the cyclical processes of PAR resulting in an enhancing change within the participant’s workplace and individual practice (Langlois et al., 2014:228). These methodological limitations will be addressed again later where I will also provide the reasons for this choice of methodology as the best fit for my workplace doctoral study.

As argued by Reason and Bradbury (2001), the characteristics of action research that are grounded in a participatory model lead to better research as the practical outcomes of the research process are grounded in the interests of those involved as opposed to an outside researcher’s agenda and interests. Perhaps one of the most defining features of participatory action research that sets it apart from other forms of social science research is the central role that non-experts play throughout the research process (Park, 2006). Sometimes, action research that is highly participatory in nature is conducted within a particular discipline in a group or is individually orientated. The individual conducting the research is often referred to as a change agent and is either collaborating within the organization or is from outside the organization. In addition, the goals and objectives of action research include methods to either improve or transform one's practice (Herr and Anderson, 2005). As action research often addresses participant concerns or problems of practical importance, data is systematically collected and findings analyzed to affect change, or at the very least modify current practice (Burns, 2005).

The participatory model supports research that is context specific and designed to involve action within one's local institution. One of the greatest benefits of action research is its practicality within any discipline where research projects are conducted in a practical setting. As described by Koshy et al. (2011), this renders action research an influential model for numerous reasons. Action research is a project that can be conducted within one's own research setting by a local platform that involves continual evaluation and modification as new theories emerge from research. Action research is also credited for its ability to promote an interest in research particularly in those who may not have previously been that involved. I find this aspect of action research extremely important to the development of this new tool, for as I previously described, the STARD tool has not been utilized to its full advantage. Therefore, one of the main goals was to develop a tool that deemed relevant and pertinent to the interpretation of radiology diagnostic accuracy trials. Conducting this action research with participants from the department will increase their level of motivation, as interpreting the literature with a tool specific to radiology will render the research project meaningful to those involved.

In summary, I am choosing participatory action research as my methodological approach as action research is concerned with scientific inquiry that is collaborative and where participants have a vested interest in the project as it is deemed applicable to practice and context specific. Radiologists continually evaluate their clinical judgments in an informal fashion with respect to their practice paradigms making recommendations for further study (Grant, 2002). In other words, based on their initial diagnostic evaluation they may suggest that further evaluation is required. Such choices are often evidence based whereby interpretation of the current literature is of paramount interest. Action research goes beyond this level of informal evaluation as collaborative processes are developed that include formal evaluation, discussion, listening, close observation and critical reflection. Concomitantly, action research cultivates reflection based on the interpretation of other's reflection and knowledge is developed through action. It is through action research that problems can be solved which may result in improved practice models due to the research findings which is not final or absolute (Koshy et al. 2011).

3.7 Methods – Phase One

Participation in this study consisted of two distinct phases. In phase one, a needs assessment was done to determine if the working group would require a revised tool specific to radiology. Pending that the radiological experts deemed a new tool beneficial, a Delphi technique was performed to study the current STARD tool. This working group consisted of 8 radiological experts from each imaging domain.
The composition of this working group was as follows:

RadSTARD Working Group

Type of Radiological Expert as per Body System
Chest
Breast
Body
Neuroradiology
Angio
Musculoskeletal
MRI Physics
Cardiac

The department of medical imaging at The Ottawa Hospital is comprised of 60 radiologists, whereby 34 are actively involved in numerous research initiatives of varying complexities. The hospital is an academic teaching center wherein research participation and publishing are heavily endorsed and respected by the department and academic officials from the University of Ottawa. This is why I chose this group of radiological experts, and a member from each body system and invited them to participate in the doctoral project. Those chosen to participate were sent a blinded email invitation that introduced them to the project with information on the overall context of the study, purpose, aims and objectives (Appendix 9). They were assured that their anonymity would be maintained throughout. This is particularly relevant in this scenario, as those who agreed to participate would also be agreeing to collaborate on a needs assessment and the Delphi technique. The purpose of the initial needs assessment was to come to a consensus on which items the radiological experts deemed were essential when interpreting the quality of diagnostic accuracy trials specific to radiology.

In a follow-up discussion with my academic advisor from Middlesex University, plus several radiologists within our department, it was recommended that I follow some of the principles utilized for the development of STARD, and another reason why I chose action research as the methodology for this study. I consider the rounds of the Delphi technique as synonymous to action research cycles. As described by Hsu and Sandford (2007), data generated from the Delphi technique can consist of both quantitative and qualitative data. In addition, action research is a research methodology that allows for flexibility in the research process so utilizing the Delphi technique in this phase of the study worked well (Koshy et al., 2011, Hsu and Sandford, 2007).

The radiologists invited to participate in the Delphi technique were sent an initial draft version of STARD-DI (Standards for Reporting Diagnostic Accuracy – Diagnostic Imaging) and an initial list of questions to determine which items they believed needed to be included in the amended STARD – DI checklist (Appendix 10). Each round of questions of the Delphi was provided to the participant via Survey Monkey questionnaire as it allowed for their anonymity. Initially, it was anticipated that it would take 4-5 rounds of the Delphi technique to finalize the items that were considered the most relevant in creating this revised checklist. Those participating were offered two weeks to respond to the questions. Reminder notices were sent out if their responses were not coming to me in time for analysis before creating the next set of questions to disseminate to the group. Once a consensus was met with the working group, and the revised checklist developed, it was piloted within our department to seek validation of the tool.

3.7.1 Action Research Cycle Spiral for Phase One – Developing the Tool

Plan – provide the participants (radiological experts) with an initially revised STARD-DI list as a starting point. Ask the experts which items they think should be included in the revised tool (Appendix 10 and 11).
Act + Observe - on the process and consequences of the changes (based on their responses).
Reflect - on these processes and consequences to develop a revised version of the new tool based on their responses and send back to the experts.

These steps were repeated until the new tool was developed.

Of the 34 radiologists actively participating and collaborating in various research initiatives, eight were invited to participate in the needs assessment and Delphi technique required to develop the revised STARD-DI checklist, subsequently renamed. The Delphi technique was completed after two rounds as consensus was met. The new tool was named RadSTARD (Radiology Standards for Reporting of Diagnostic Accuracy Studies).

3.8 Phase Two

Once the tool was developed, the second phase was to validate it by testing the tool between the residents and Fellows. Initially, I had planned to ask the residents and Fellows to compare the new tool (RadSTARD) to the STARD tool after they read 2 radiology diagnostic accuracy articles. However, given that the basic premise of the STARD tool is to use as a guide when reporting the results of a diagnostic accuracy article, it did not seem appropriate to ask the residents and Fellows to make a comparison between the two tools. I was also concerned that they were not all that familiar with STARD, and if asked to compare it to RadSTARD, it could lead to confusion. Instead, the residents and Fellows were asked to rate their confidence level when interpreting one radiology diagnostic accuracy study that was specifically chosen as it included items that could be found within the newly revised tool called RadSTARD (Henes et al., 2012).

In addition to developing the RadSTARD, it was essential to also develop an elaboration document as this would describe what each component of the tool is measuring when interpreting the literature (Simera et al., 2008). Hence, an explanatory document was developed once the RadSTARD was finalized. In addition, a 2-page summary of RadSTARD was also developed which provided a brief summary of what each item was measuring or describing. Concomitantly, each item of STARD that correlates to each item of RadSTARD was listed as well. As described by Bossuyt (2014), the provision of an explanatory document aids the reader in facilitating the use of a checklist, plus it aids comprehension of the results as reported within the article. Each item within the RadSTARD checklist was described allowing for additional interpretation, clarification of the item under evaluation, and examples of biases that could impact overall applicability were all illustrated. This document was provided to the residents and Fellows as a user guide as they referred to the new tool to rate their confidence level as they interpreted the radiology diagnostic article. They were also provided a second diagnostic accuracy article that was reported according to the STARD tool (McComiskey et al., 2012). Hence, the STARD checklist (Appendix 12) and elaboration document (Bossuyt et al., 2003a) was provided for comparison to the newly developed RadSTARD tool and elaboration document. They were asked to review both articles (McComiskey et al., 2012, Henes et al., 2012) but only rate their level of confidence when interpreting one radiology diagnostic accuracy article (Henes et al., 2012).

Within this department there are approximately 31 residents and 17 Fellows whereby academic teaching lessons and research rounds are routinely held. I contacted the radiologists in charge of academic rounds for the radiology residents and Fellows seeking their permission to present my study at academic rounds, explaining the purpose and goals of my study and clearly defining what the role of the participants would be. Permission was granted for me to present to the residents and Fellows and to invite them to participate in my study (Appendix 13 and 14). These sessions were conducted with the residents and the Fellows at separate times. Each group was provided ample time to ask any questions so that they clearly understood their involvement and to ensure that any of their concerns were alleviated. Due to the nature of action research, sometimes it is difficult to anticipate everything that could occur until the project unfolds (Herr and Anderson, 2005). Suffice it to say, all concerns raised were addressed whereby the flexibility and awareness of any conflicting commitments were acknowledged and managed on an individual basis. The only concern raised was the time it would take to complete the survey.

My methods for piloting RadSTARD at Academic Rounds include the following cycles:

Cycle 1:

- Provided a power-presentation (Appendix 15) that included the following information:

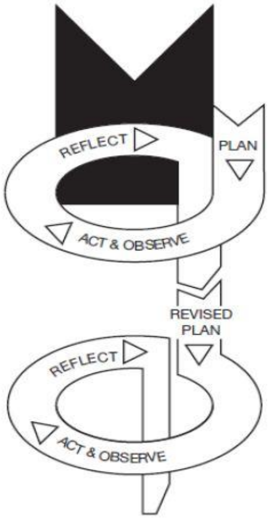
- Brief introduction to the doctoral study
- The purpose – aims of the research (to develop a reporting tool that is specific to radiology diagnostic accuracy studies)
- How the tool was developed by radiologists for radiologists via a needs assessment and Delphi technique
- Listed the differences between RadSTARD and STARD
- Explained that an e-mail would follow with the link to 2 radiology diagnostic accuracy trials for them to read. Asked them to refer to the RadSTARD tool as they read in order to determine if the tool increased their level of interpretation after they read a radiology diagnostic accuracy article. They received a second article reported as per the STARD recommendations to refer to
- The RadSTARD elaboration document, plus a 2-page RadSTARD summary was provided to participants to provide details pertaining to each item of the RadSTARD
- They also received a copy of the STARD tool, plus elaboration document for reference only to compare to RadSTARD
- Their responses were recorded via Survey Monkey, which allowed for anonymity, plus this data capture would later provide data analysis
- After they read the diagnostic accuracy article, they were asked to rate their level of confidence when interpreting each item of RadSTARD via a Likert Rating Scale where 1 is least likely – 10 is most likely
- Opened the floor to questions and addressed any concerns

Cycle 2:

Met with the residents 3-4 weeks later to present the results from the Survey Monkey Questionnaire

- Provided a second power point presentation (Appendix 16)
- Answered any questions
- Asked them to complete a short survey (via pen and paper) on the Usefulness of RadSTARD and drop it in a box in the teaching classroom before they left Academic half day
- If they did not complete reading the articles and answer the initial Survey Monkey questionnaire, they were exempt from completing the questionnaire on the Usefulness of the RadSTARD

Although I had planned on presenting my study and inviting the residents and Fellows to pilot the RadSTARD at their academic half days, not all physicians could attend the rounds as scheduled. Hence, I met with them on numerous occasions throughout the ensuing weeks in order to see all the Fellows and residents and invite them to participate. The results from the two phases of testing (Survey Monkey questionnaire and follow-up pen and paper questionnaire on the ‘Usefulness of RadSTARD) was done with the purpose of collecting mixed data in keeping with one of the basic tenants of action research, triangulation (Koshy et al., 2011).



Kemmis and McTaggart's action research spiral (Koshy et al., 2011).

The action research cycle spiral above for Phase 2 consisted of the following steps:

Plan – provide the participants (radiology residents and Fellows) two radiology diagnostic accuracy articles to review along with the RadSTARD, RadSTARD elaboration document and 2-page RadSTARD summary document. I provided the STARD and the STARD elaboration document as well, but as a reference document only. The participants were asked to rate their level of confidence when interpreting the results of one article that was written with items that could be found with the RadSTARD (Henes et al., 2012). The goal was to determine if these items increased the participant's level of confidence when interpreting the article.

Act - Of those who agreed to participate, once they had read the two articles and reviewed the accompanying documents, they rated their level of confidence in interpreting the article by Henes et al. (2012) by completing the Survey Monkey questionnaire.

Observe - Responded to any queries from the participants and recorded any of their feedback as it was provided to me.

Reflect - on these processes and consequences when analysing the data to present as descriptive data when conveying findings of the interim analysis.

Once the data was analysed the results were provided to the residents and the Fellows and the action research spiral was completed again. The next validation phase consisted of them completing a one-page pen and paper questionnaire on the 'Usefulness of the RadSTARD' that allowed for qualitative responses as well.

3.9 Data Capture and Analysis

Throughout this study there have been several distinct phases where data was captured and analysed. These include the initial needs assessment done by the radiological experts who agreed to participate in this study. The items deemed essential when interpreting reporting of diagnostic accuracy studies specific to radiology where analysed and then provided back to the radiological experts for their review and eventual consensus on which items they deemed should remain in the final RadSTARD tool. This was done via the Delphi technique.

As the overall aim of action research is to create change, evaluation includes reflecting on the impact of my study (Gray, 2013). The new evidence created was the RadSTARD tool. As with any reporting guideline, an accompanying elaboration document was also written. Given the magnitude of the elaboration document, a two-page summary of the RadSTARD tool was also written. In an effort to validate the new evidence from the RadSTARD tool it needed to be judged by others. Henceforth, the next phase consisted of validating the RadSTARD tool with the residents and Fellows from the radiology department.

As action research is about learning, inviting the residents and Fellows to participate in the validation phase of evaluating the RadSTARD at their academic half-day sessions presented the perfect milieu. Again, this phase of the study was also comprised of two distinct phases.

Phase 1: At different instances, a PowerPoint presentation was provided to the residents and Fellows, which described the purpose of the study and what their participation would entail. Following my presentation, the residents and fellows were e-mailed the consent and accompanying documents (Appendix 17). These documents included two radiology diagnostic accuracy studies whereby one article had been published according to the STARD recommendations and the other article had not. They were asked to review both of these articles and rate their level of confidence for interpreting the article by (Henes et al., 2012) in reference to the RadSTARD tool. In order to accomplish this task they were provided the RadSTARD tool, elaboration document and two-page summary of the RadSTARD tool. They were also provided the original STARD tool and the elaboration document for the reference only.

Several methodologists were consulted within my institution on this project with respect to data capture and analysis. With respect to assessing their overall interpretation of the article and confidence level by the residents and Fellows, their responses were obtained from a Survey Monkey questionnaire whereby each item of RadSTARD was individually rated from 1 (least confident) – 10 (most confident) (Appendix 18). Providing 10 variables enabled capturing and coding of the data to conduct a chi-square test to identify trends between the two groups with respect to confidence levels when interpreting the literature using the RadSTARD. These results could also determine if the revised tool resulted in any significant difference in homogeneity between the groups (Dawson, 2008).

Phase 2: The next stage also involved providing the residents and Fellows a subsequent PowerPoint presentation on the interim analysis of the data that was collected from those who participated in the previous stage. The purpose of this stage was to obtain the residents and Fellows opinion by completing a questionnaire on the ‘Usefulness of the RadSTARD’ tool as they now had experience utilizing it. Developing a method to collect this additional data was challenging, as I was unaware of any standardized questionnaires to assess for the usefulness of a reporting tool. This is a major limiting factor of the majority of these tools. The thrust is on publishing, with little to no validation done in advance (Moher et al., 2014b).

Therefore, a questionnaire was developed to assess the ‘Usefulness of the RadSTARD’ tool (Appendix 19) by modifying a questionnaire published previously and used to assess the usefulness of QUADAS (a tool that measures the quality assessment of studies of diagnostic accuracy included in systematic reviews). Some of the items were not pertinent to this study as the RadSTARD is not a tool to be used for systematic reviews (Whiting et al., 2003). Although there was no previously validated questionnaire to refer to as a template when developing the questionnaires for my study, certain principles of questionnaire development were followed.

For example, the questions were simple and open-ended. In advance to utilizing the questionnaires they were piloted to the senior radiology resident and one Fellow to determine if they thought revisions to the questionnaires were required. Although concern of time to compete the study tasks was expressed, the content of the Survey Monkey questionnaire and ‘Usefulness of the RadSTARD” questionnaire were not challenged. The questionnaires were not modified based on the feedback; however, methods of validation of the revised tool were modified as described earlier (Koshy et al., 2011). All documents were reviewed and approved by Middlesex University and my local regulatory board prior to usage in the study.

As it is imperative that triangulation of data be accomplished when conducting action research, this was achieved by analyzing the quantitative data from the Survey Monkey questionnaire, plus the residents' and Fellows' responses to the 'Usefulness of the RadSTARD' questionnaire which consisted of open-ended questions as well as space for them to insert their own comments. Henceforth, mixed methods were used as both quantitative and qualitative data were collected and analyzed.

Additional qualitative data was documented in my reflective journal, which was quite beneficial in writing up the results as it allowed me to find my voice amongst the data. Reflecting on the entire process also contributed to my development as a professional researcher (Koshy et al., 2011); therefore, all conversations and observations were recorded. My critical friends, mentor and consultant were kept apprised of my study progress throughout all stages and helped navigate through challenging aspects of my methods with strategies that allowed me to overcome various situations (Herr and Anderson, 2005).

One concern of action research is that the data generated is not considered generalizable. One way to address this concern is by declaring that the overall goal of action research is to generate knowledge not generalizability (Koshy et al., 2011). Hence, the different phases of this study have been presented in poster format for two consecutive years at the Middlesex University's summer research conference, and abstracts published in a work-based learning journal. An abstract was submitted to RSNA (Radiological Society of North America) and was not accepted for the 2015 annual conference. However, I am confident it will be accepted later as data will be available allowing for recommendations specific to the field of radiology diagnostic accuracy research. Dissemination of these findings is important as it will illustrate to those involved in research how they could utilize similar methods in their own settings in an attempt to replicate the study or simply to validate a similar document (Koshy et al., 2011).

3.10 Reflections on the Overall Response Rate

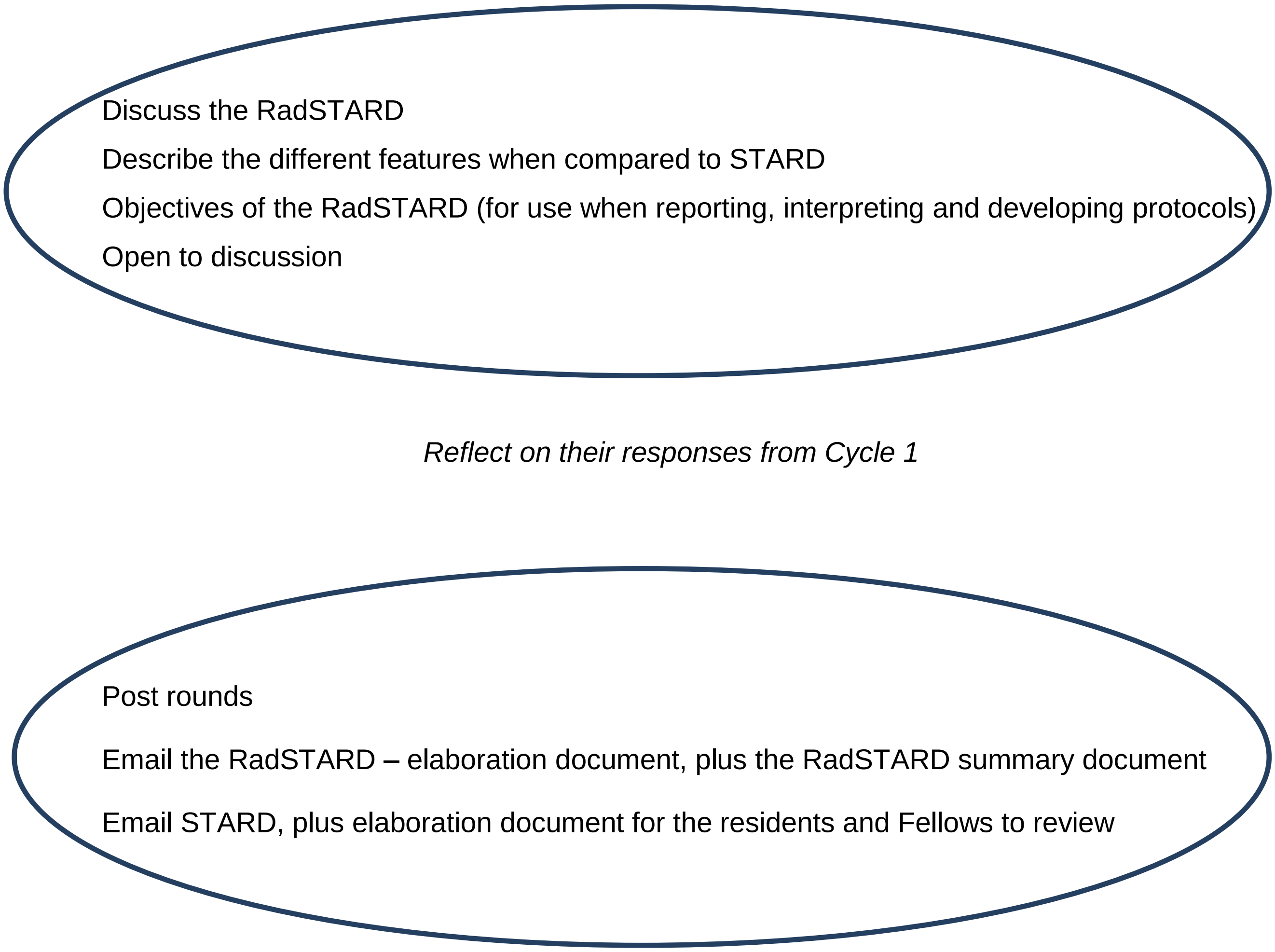
Throughout the conduct of this study there were several phases that were particularly challenging. I was fortunate, given my position in the department, to have the opportunity to readily discuss my ideas with various radiologists although not involved in the study, were well aware of it and were very supportive with their recommendations. I am referring to the validation phase of the RadSTARD tool. The biggest question that daunted me was how I could obtain the residents and Fellows opinions of the RadSTARD tool and accompanying document with a good response rate whereby the methods to attain their opinions would not be deemed as coercive.

Initially, it was suggested that perhaps we could review the articles at Journal Club and then the residents and Fellows could answer questions on rating their level of incompetence for interpreting the articles with the RadSTARD via the clicker response. This clicker response method would have been a way of electronically capturing the respondent's opinion. When I reflected on this option, I thought this action would be problematic in that the residents and Fellows would not have enough time to think about what was being asked of them, and therefore, I decided not to use this method.

So, I went back to the drawing board and strongly considered the action research spirals. As I reflected on what the overall goal of validation was I referred to the action research cycles by Kemmis and McTaggart whereby the following action research spirals were developed:

Figure 1 - Present the RadSTARD to Residents and Fellows at Academic ½ Day

Cycle 1 *Planning a Change*



Discuss the RadSTARD

Describe the different features when compared to STARD

Objectives of the RadSTARD (for use when reporting, interpreting and developing protocols)

Open to discussion

Reflect on their responses from Cycle 1

Post rounds

Email the RadSTARD – elaboration document, plus the RadSTARD summary document

Email STARD, plus elaboration document for the residents and Fellows to review

Act (sending email) & Observe

Reflect (responses to cohort)

Reflect on the process/method of delivery

Email link to radiology diagnostic accuracy trial

Ask residents/Fellows to rate their responses on the RadSTARD summary (Likert rating)

Reflect on the process and then plan to meet

Meet to discuss in focus group

Discuss their thoughts on the utility of the tool

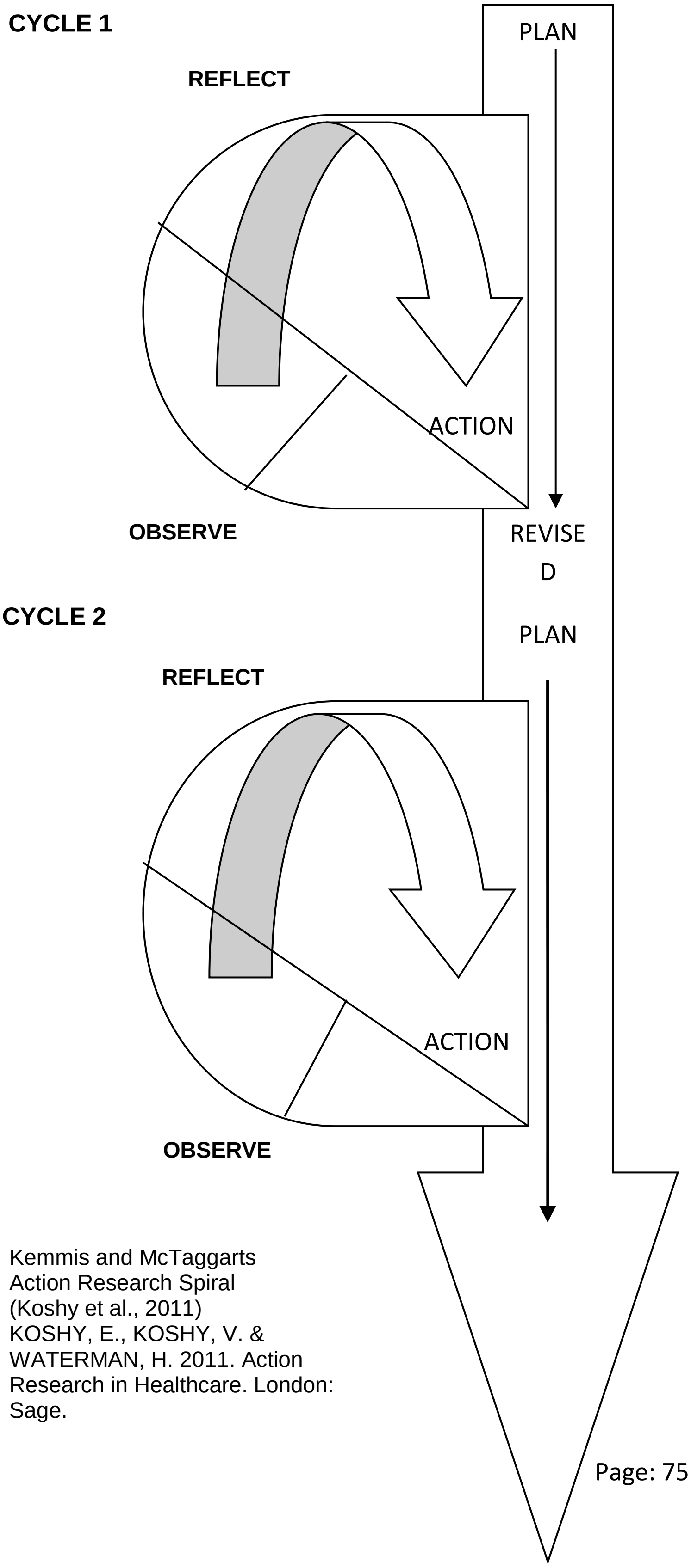
Would it change their research practices

Is it useful?

Gather their responses to reflect upon them and the overall outcome.

Kemmis and McTaggart (2000) in (Koshy et al., 2011)

Figure 2: ACTION RESEARCH SPIRALS for Phase 1 and Phase 2



Phase 1

Plan: Provide radiological experts with revised STARD_DI.
Act: Ask experts which items should remain
Observe: The processes and the change based on their recommendations
Reflect: On processes and changes to develop a revised version of the new tool based on their responses and send back to the experts

These steps are repeated until new tool is developed

Phase 2

Plan: Provide residents and Fellows with 2 radiology articles to review with RadSTARD and STARD. Asked to rate their level of confidence with RadSTARD only.
Act: Level of confidence is rated via Survey Monkey
Observe: Respond to any queries and record feedback.
Reflect: These processes and consequences when analysing the data to present as descriptive data when conveying findings of the interim analysis.

Once the data was analysed the results were provided to the residents and the Fellows and the action research spiral was completed again.

Kemmis and McTaggarts Action Research Spiral (Koshy et al., 2011)
KOSHY, E., KOSHY, V. & WATERMAN, H. 2011. Action Research in Healthcare. London: Sage.

3.11 Limitations of my Methods

Although on-line questionnaires such as Survey Monkey are convenient, easy-to-use and can perform data analysis they are not useful if you don't have data. For instance, a physician may agree that they will complete the survey but given the anonymity in their responses I wasn't sure they had completed the survey. The Chief resident and the Uber Fellow suggested sending out reminder e-mails to the residents and Fellows as they have a very heavy academic load in their training (Appendix 20). The work schedule of physicians and medical trainees is very demanding, so participating in a survey for my doctoral study may not have been a high priority (Flanigan et al., 2008).

When I gave my departmental power-point presentations, I realized that several physicians for both groups were not present. They were absent for any number of reasons including the possibility that were post-call and were at home. Therefore, I offered to meet with those physicians on an individual basis and I presented my power-point presentation again explaining what was being asked of them. I found that this one-on-one instruction was readily received, and as I reflected on my reflection-in-action, I realized how limited their experience was when using these reporting tools.

By merely walking through the corridors within the department and seeing various residents and Fellows was also a reminder to them if they had not completed the survey. Indeed, when they saw me some of them would say, ‘I have not completed your survey yet but I will do it this weekend’. The residents and Fellows were given a projected end-date to participate in Phase 1 which was reading the articles and completing the online Survey Monkey questionnaire. This date was negotiated with both working groups, which I thought was important as it made them feel like true collaborators in this action research study. As the days progressed to weeks, I was able to monitor how many had participated in the survey and as the time drew nearer to the projected deadline I became anxious in the lack of response rate. Therefore, I elected to individually e-mail each participant from the resident and Fellow working group. This action was highly successful as evidenced by the rate of response.

Of the 17 Fellows, 15 of them participated in the on-line Survey Monkey questionnaire for a responder rate at 88%. And from the 31 radiology resident group, 18 residents responded resulting in a responder rate of 58%. In general, the responder rate for completing surveys amongst physician groups has been reported as less than those from the general population. As described by Flanigan et al. (2008:4137), the mean rate of response for physician groups was “54% compared to 68% mean response rate among non-physicians”.

3.12 How to Measure Change from Action Research and Issues of Trustworthiness

Given that the basic premise of action research is to evoke change within one's practice, I considered the final aspect of data collection in my study to be the most important one (Gray, 2013). The final questionnaire on the ‘Usefulness of the RadSTARD’ tool will result in collecting subjective data from those who agree to participate in this final phase of the study. There are pros and cons to every method utilized in research. As described earlier from my research experience in radiology and from my literature review, it was clearly evident that the STARD tool was not being used to its full potential in the reporting of radiology research. Concomitantly, I also knew that the residents and Fellows were not using the STARD tool. Therefore, this presented challenging aspects in the validation phase of the tool. I am fully aware that when conducting action research it is highly desirable to conduct interviews with the working group collaborating on the study (Gray, 2013). In that there was certainly a proportion of the physician group that were not that comfortable in working with reporting tools such as the STARD tool, I needed to devise a way that would allow for validation of the RadSTARD tool whereby the participants were given enough time to really consider what was being asked of them, what they were learning, and how to respond accordingly to the question at hand. I think this was most beneficial for them as participants/collaborators of this research, which resulted in enhancing the trustworthiness of their responses. In addition, given my tenure I have attended many academic rounds and found that by and large when the physicians were asked to respond to a question within a group there is often dead silence. No one would want to be pointed out and given their lack of exposure with the reporting tool I did not want to make anyone feeling uncomfortable. My goal was to enhance the response rate, not hinder it.

3.13 Ethical Considerations

Although this project did not involve consenting patients, children or vulnerable populations ethical ramifications were anticipated in the planning phase of how I would conduct my work-based research. These ramifications were slightly enhanced due to the evolution of action research (Koshy et al., 2011). Fox et al. (2007) claims that all research is prone to bias, especially when research is conducted by an insider researcher within their own place of work. This issue can be by-passed by conducting the research elsewhere but in this case it was not possible to conduct my research at an alternative location. As insider researchers, Costley (2006) argues that the knowledge insiders have of systems, as well as those designated to the project may render an ethical situation as opposed to those who can conduct their research and then leave the research space. This ‘ethics of care’ is noteworthy given my work-based project took place within our department which is also my workplace (Costley, 2006). However, as described in my DPS 4520 paper, my extensive experiential learning in clinical trials coupled with my graduate level education, provided me with the additional knowledge and skills required to manage the research program and confidently collaborate with all involved in the research. As an experienced and ethically responsible research professional, I was fully aware of the ethical implications that needed to be maintained throughout the research process, from deciding on the purposes and aims of the project, to dissemination of findings.

Collectively, I worked for the radiological experts who were invited to collaborate on the development this tool, later piloted within the department. As noted by Meyer (2000), this close affiliation can enhance the significance of the research process making it more meaningful to those involved. I would agree with this statement. As participation is a key proponent of action research, the challenge was meeting this caveat; actively participating was part of the change process. As the level of willingness by the physicians to participate in my action research study was a commitment that shouldn't be taken lightly, the consent form clearly described what their involvement would entail; however, not all research processes are known a priori, which is congruent with action research as it could theoretically impact those who participate (Herr and Anderson, 2005). The following section will review some of the key ethical factors I considered with respect to my study.

3.14 Consenting

As previously described, this study was conducted in two parts. The radiological experts chosen to develop the revised STARD tool by completing a needs assessment and then participate in a Delphi technique was done via email invitation (Appendix 9 and 11). Only one physician decided not to participate and this decision was respected. Therefore, another radiological expert took their place. When conducting a Delphi technique, it is paramount that the correct participants are chosen as the results generated reflect on the group (Hsu and Sandford, 2007). In order to cover all aspects of radiological imaging and body systems, a radiologist or PhD scientist participated were invited to participate in the research.

Although they were provided a Participant Information Sheet and Consent (Appendix 9) to review in advance, their signature to participate was not required as agreeing to participate was considered implied consent. Alternatively, if an individual decided not to participate their decision was respected.

Given my role within the department, others may view me as a person in a position of power. As an insider action researcher all efforts were made to avoid potential participants feeling coerced to participate by asking them which way they would like to participate and freely consent (Herr & Anderson, 2005). It was beneficial to all to discuss my methods with the chief resident and one of the senior Fellows in advance. As discussed earlier, action research is often conducted with a group as a whole and with those looking to implement a change within their department. Although they were not initially involved in the protocol development phase, their participation later on was integral to the success in the learning that occurred as they were given the opportunity to reflect on the merits of the RadSTARD tool, which in turn could bring about a practice change. Due to the cyclical nature of action research, it is also recommended that "processual consent" continually be sought throughout the project in addition to original consent as next steps are explained to the participants (Herr & Anderson, 2005:120). Hence, several regulatory approved consents (Appendix 21) were provided with each phase of this study.

3.15 Chapter Summary

In summary, as I recount the methods of my doctoral study, this in-depth level of reflection greatly enhanced my ontological and epistemological stance and rendered me clarity with respect to research paradigms. On further critical reflection, I also realize how my positionality fit with action research methodology, and in particular participatory action research, as I was accustomed to working with research initiatives that prescribed to a similar model. In particular, I am referring to the participatory nature of action research essential to all phases of this study. As interim results were disseminated, they also raised awareness as to the importance of quality reporting when interpreting the results of diagnostic accuracy results. Although this project is specific to developing guidelines for radiology, it is hoped that in time, and with publication, similar tools will be developed in other areas of medicine that conduct and report diagnostic accuracy studies. The STARD tool was published to be used when interpreting all diagnostic accuracy trials (Bossuyt et al., 2003b). I would argue that the current tool is not a 'one model fits all' tool and that there is a gap in knowledge that required further study. At least in radiology, the need was there to study this tool with robust methodology and rigor in a participatory fashion that action research provides. Upon dissemination of my final results to the current body of literature available on reporting guidelines, I will illustrate why the need was there to examine the current tool in such detail. Concomitantly, I will describe how participatory action research methodology allowed for critical reflection that was required to study the current tool to determine if STARD should be revised and if these revisions were deemed more beneficial when interpreting the quality of methodology as reported in radiology diagnostic accuracy studies.

4. Chapter 4 Project Activity

4. 1 Conduct of Action Research

The action research study was undertaken within the radiology department at The Ottawa Hospital, Ontario, Canada. Within the health sciences, action research is a methodology well suited to the goal of collaboratively improving practice (Koshy et al., 2011). Action research often involves utilizing qualitative data collection methods such as conducting interviews and observation (Meyer, 2000). One of the most challenging components of the study was determining how to effectively conduct the action research study so that the methods were congruent with action research practices, plus met with scientific validity and rigor. As described by Cook (2009), one of the main purposes of action research is to develop rigor from the mess that occurs while utilizing the methodology. Arguably, my past research knowledge and professional experience gave me both the ability to conduct the research with authority, and the insight required to critically reflect on my actions at every step throughout the project. As I had to make challenging decisions based on my keen awareness of the clinical research arena in which I was planning on conducting my doctoral study, at times my research actions became messy. The goal of this action research project was to develop a reporting tool for radiologists to use for reporting and interpreting radiology diagnostic accuracy studies. The two distinct phases to this study were as follows:

4.1.1 Phase 1: Development of revised diagnostic accuracy-reporting tool specific to radiology.

Initially, eight radiological experts were asked to participate in a needs assessment and Delphi technique required to develop the revised reporting tool specific to radiology STARD_DI (Standards for Reporting Diagnostic Accuracy – Diagnostic Imaging). This draft tool was called the STARD_DI later renamed as described below.

4.1.2 Phase 2: Validation of new tool.

Once developed, the new tool was initially named the RADART (Radiology Diagnostic Accuracy Reporting Tool). However, I later elected to amend the name of the new tool to RadSTARD (Radiology Standards for the Reporting of Diagnostic Accuracy Studies), as the newly developed tool was comprised of STARD items, plus items specific to radiology diagnostic accuracy studies. Piloting of the RadSTARD was done with the radiology residents and Fellows as a form of validation to determine its relevance and measure outcomes.

4.2 Phase 1 Activity

Methods for studying the STARD tool have been described in the methods chapter. Since this was an action research study, to develop and validate the tool I collaborated not only with my department, but also with my mentor (radiologist) and two critical friends. Both of these physicians were experienced researchers, knowledgeable on research methods such as the Delphi technique, and both were familiar with the STARD tool.

Principally, I did the initial development of my doctoral study proposal, although I discussed many ideas with my mentor. In addition, my advisor from Middlesex University was also informed of my actions throughout the study. As my department had partially funded my studies, they were also keenly interested in what I was studying, and how I was planning on conducting the research. It was suggested that I discuss my proposal with one of the radiologists with an interest in diagnostic accuracy research. This physician was initially quite interested in my research proposal and thought I could merely replicate what other researchers had done when previously published reporting tools were amended or revised. By way of example, I am referring to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), developed as a critical appraisal tool when QUOROM (Quality of Reporting of Meta-analyses) standards had become out-dated. However, this was not the purpose of my study. My plan was to study the current STARD tool with the goal of amending it to make it more specific to radiologists when reporting and interpreting diagnostic accuracy research specific to radiology.

Interestingly enough, a few weeks later the same physician attended the Cochrane colloquium and from that meeting sent me an e-mail stating that QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies) had just been presented. He tried to discourage me from conducting my research suggesting that since QUADAS-2 was already developed there was no reason or need for me to continue on with my proposed research. I was at first taken aback by his correspondence, however; I explained to him that the QUADAS-2 document was to be used for meta-analyses while the STARD checklist was not developed for this purpose. Certainly, many researchers have used the STARD tool to conduct systematic reviews, but it was not developed with the purpose of conducting meta-analyses. When I explained this to the radiologist he concurred and added that my study was a great idea.

4.3 Research Ethics Board

As a senior research professional, I am very well versed with the process of obtaining research ethics approval in advance to conducting any research. I am also aware of the need for apprising the research board with any amendments to a research proposal, and providing supporting documents throughout the conduct a study. Once I received ethical approval from Middlesex University (Appendix 22), I applied for regulatory approval from my local ethics board (Appendix 21). This was a long and arduous task. The research board is extremely thorough and given the volume of research conducted at the University, it sometimes took many weeks to respond to submissions and/or amendments. It was also required that I obtain regulatory approval in two separate phases. The first submission to the research ethics board was done to request approval to conduct my study. Only after this phase was completed could I submit an Amendment to the research ethics board requesting a second approval. Within this Amendment, I supplied the research ethics board the newly developed tool (RadSTARD) and elaboration document describing what each item of the tool was measuring. As the elaboration document was quite lengthy, I also submitted a 2-page summary document describing each item of the RadSTARD. In addition, I also supplied all regulatory documents such as my consent forms, and the questions I would be asking of the cohort who willingly participated (Appendix 17). It took many months to establish regulatory approval for both of these phases. Ethics approval was also required when I changed the name of the revised tool, keeping in mind that the draft tool was initially called STARD_DI, then RADART and finally RadSTARD (Appendix 21). Renewing the study annually was also required from the local ethics board (Appendix 21). Throughout all of these phases, my Middlesex University advisor was aware of any proposed changes to my original doctoral project and further regulatory approval from the University was not required as the changes were not substantial.

Based on the initial review of my application, two very interesting questions came back from the local research ethics board requiring my response. The first question was regarding intellectual property (IP) and contained two parts:

a. Have you made yourself aware of intellectual property issues?
b. Have you clarified with the participants the ownership of data?

Fortunately, I was able to consult with an expert within our institution on intellectual property who assured me that OHRI (Ottawa Hospital Research Institute) was of the opinion that any intellectual property developed as a result of a doctoral study did not fall within the scope of intellectual property (IP) to which OHRI retains ownership by policy. Therefore, they claimed no ownership of such IP, and as far as they were concerned, that were no IP issues (Appendix 23).

The second and arguably more pressing question from the research ethics board was whether the current STARD tool was copyrighted? In order to answer this question I needed to correspond with Dr. Patrick Bossuyt, the leading author of the STARD tool (Bossuyt et al., 2003b). He initially responded quite quickly stating that I was free to use the STARD tool unmodified. When I explained that my study would involve studying the current STARD tool to determine if amendments were required specific to radiology, he expressed concern that there would be confusion around the current STARD tool. Dr. Bossuyt is of the belief that the current STARD tool was developed for reporting diagnostic accuracy studies within any area of medicine. However, based on my personal independent study and knowledge, as well as the opinion of others within the literature (Wilczynski, 2008, Ochodo et al., 2013), the current STARD tool was not being utilized to its full potential, at least not in the reporting of radiology diagnostic accuracy studies (Saba et al., 2012, Saba et al., 2011, Malcius et al., 2009, Bardou et al., 2013, Crim et al., 2013, Georgantopoulou et al., 2008, Chen et al., 2015b, McComiskey et al., 2012, Laméris et al., 2009, h-Ici et al., 2012, Tseng et al., 2015, Fratz et al., 2009, Miller et al., 2009).

Fortunately, when the purpose and methodology of my study were further described to Dr. Bossuyt, he confirmed that the copyright for the STARD checklist had been waived and that I was free to create a new tool (Appendix 24). He also added that once developed, perhaps some of my findings might be relevant to future versions of the STARD tool. The STARD revised list was recently published in October 2015 and is discussed further in the literature review chapter.

4.4 The STARD_DI Working Group

The development of the STARD_DI working group was carefully chosen in collaboration with my mentor. These eight radiological experts specialize in one particular imaging domain, for example cardiac or breast imaging. As there are eight domains in radiological imaging, I chose eight radiologists to participate in the study. Of the eight radiological experts, an MRI physicist with a PhD in MRI physics and very well versed in research was also invited. Upon initial invitation, seven out of the eight radiological experts agreed to participate in the study consisting of a needs assessment and the Delphi technique in the creation of a new tool. On reflection, I remember how perplexed as I was with the delay in response to my initial invitation, especially since I had over the years worked so closely on previous projects with most of them. This was my first learning lesson as a researcher conducting a study within my own department. As the principal investigator, I was the one now leading the study, and reversing the roles. On further reflection on my reflection-in-action, when I called each physician to inquire if they had any questions about their involvement this helped them to better understand what their role in the study would entail (Schön, 1987a). As these were the initial steps of conducting my own study, I recall discussing with my mentor how different it was to be the one actually conducting the study rather than providing the administrative support for somebody else’s study.

The radiologist initially chosen as the eighth member of this collaborative working group stalled her decision to participate for many weeks. Although some of the other radiologists also expressed concern about the time commitment that it would take to actively participate in the study, this particular physician was also concerned about being included in future publications resulting from this study. After several email exchanges inquiring as to her decision to participate, by chance I finally met her in the hospital library. When politely asked if she had made a decision, she again expressed time commitment concerns. I explained that I would send out an initial list (STARD_DI) to the radiological experts asking their opinion of what they thought should be maintained on the list. Once their responses were obtained, I would then ask for their opinion on a revised list based on the answers, and this would continue until consensus had been met, and the revised tool created. She had no difficulty understanding what would be involved; however, I sensed that this was not something she wanted to do and I immediately excused her from participating when she referred to my study as the “Betty Anne's tool”. I recognized that this unfavorable response was a problem of micro-political “power-sharing” and immediately decided to excuse her from any further involvement to the trial and instead inviting another expert to participate (Smith et al., 2010:409).

I found this experience quite difficult, as I had worked with this particular physician on numerous studies in the past including research conducted during her Fellowship training. Nonetheless, after relaying the outcome of this interaction with my mentor, I invited an alternate radiologist to participate and she accepted.

4.5 Needs Assessment

Enhancing one’s professional development typically requires additional learning to successfully render a change in practice (Grant, 2002). However, prior to the initial step of learning, conducting a needs assessment is an enriching endeavor whereby one’s existing epistemology is reflected upon as one links their professional practice and tacit knowledge to identify their current needs pending the topic under study. Certainly within the professional domain of medicine, needs assessments are routinely performed to identify the additional learning needs required to enhance one’s professional practice, whether formally or informally (Grant, 2002). Therefore, a needs assessment was completed prior to conducting the initial iteration of the Delphi technique, as the needs assessment provided an opportunity to affirm the radiological experts agree there was the need for a modified tool (STARD_DI).

As described by (Grant, 2002:157), needs assessments are categorized into seven main groups. In particular, one group requires “reflection on action and reflection in action”, proving most beneficial, as the needs assessment in this study functioned as the baseline for the radiological expert’s epistemological stance and research mindedness with respect to diagnostic accuracy research. In this instance, the main purpose of performing the needs assessment was to identify which items the radiological experts deemed essential when interpreting the quality of diagnostic accuracy trials as reported specific to radiology.

When I originally developed my doctoral proposal, I thought conducting a needs assessment would be straightforward. I searched for articles on learning and needs assessments, trying to find a way to conduct a needs assessment with diagnostic accuracy. This led me to question whether I should ask the radiologist if completely new items needed to be included in the needs assessment itself. For example, I contemplated adding receiver operator characteristic (ROC) curves versus conventional sensitivity and specificity analysis which were fundamental to diagnostic accuracy research (Van Erkel and Peter, 1998). On further reading, I explored the possibility of adding likelihood ratios to the list of potential items required, as the provision of likelihood ratios enables the interpretation of tests more readily by clinicians (Florkowski, 2008).

I continued reading articles on needs assessments (Grant, 2002) while concomitantly reading action research literature (McNiff, 2013). As I continued to read, I became over-whelmed with the volume of information and choices on how to conduct my needs assessment. The question daunting me was: *'how am I going to conduct an effective needs assessment?'.* I did not want to burden the radiological experts with tasks they might deem to be of little value. As I reflected on what I read and was aiming to do, I realized that it may be best if I sent them a list of items from STARD and TiDier (Template for Intervention Description and Replication) checklists and inquire which items they thought needed to be included in the needs assessment (Appendix 25). Many months ago when I was first developing my research proposal, a Senior Methodologist from the hospital suggested this idea to me. My mentor also agreed with this approach.

As described by McNiff (2013), knowledge is generated by action researchers by something they do, and as such, is a living process. Knowledge creation is a continual developmental process of new understandings, whereby it is never static or complete. Given that reality is an unpredictable and evolving process, there are no fixed answers in the creation of knowledge as answers can transform into new questions. Therefore, action researchers are constantly asking questions such as “I wonder what would happen if...” as their overall goal is to disrupt systems of knowing that are fixed as opposed to maintaining the status quo (McNiff, 2013:18).

During this initial stage of my action research, as I reflected on my prior knowledge on the use of reporting tools such as STARD within our research program, this reflection prompted my decisions for my actions (McNiff, 2013). I saw the value in reporting tools such as STARD, but my goal was how to improve it and make it more appropriate for use in reporting and interpreting of radiological diagnostic accuracy studies.

Once the needs assessment was complete, this knowledge identified some of the key items to be included in developing a reporting tool specific to radiology diagnostic accuracy. Based on this knowledge, and upon reflection, it led my actions to the next stage of my doctoral study, which involved conducting the Delphi technique with the radiological experts. At this stage I met one of my critical friends.

4.6 Critical Friend # 1

When I first began my study, I had only one critical friend; the second critical friend agreed to join in collaboration later on. Critical friend number one was a physician with an MSc in clinical epidemiology and well versed in research processes, diagnostic accuracy and the purpose of reporting tools such as STARD. As he worked at The Ottawa Hospital, this close proximity afforded me the opportunity to meet with him on a regular basis initially, with further communication conducted via email. This particular critical friend was also a colleague of my mentor, although their area of specialty was different.

The following is a summary of our meetings, plus e-mail conversations with further details provided in the appendices (Appendix 26). Our first meeting was conducted on July 14, 2014 and entailed a discussion as to which items included in my needs assessment render an opinion. At this time, not all participants had finished the needs assessment. Based on the fact that those who had participated thus far considered most of the items included in the needs assessment favourable, my critical friend wondered if some of the items might not be relevant to radiology. He suggested that I make the list for the first round of the Delphi in a fashion more succinct to interpreting diagnostic accuracy studies. For example, the first item of the STARD list states that ‘sensitivity and specificity’ should be in the title of a diagnostic accuracy study. My critical friend suggested that I provide an example of a study title with the word ‘sensitivity and specificity’ in the title to see if it should be concluded. In addition, he wondered if asking the research question was redundant and suggested ways to rank order my list of items so that there was some semblance of order.

The end of July 2014 marked the completion of the needs assessment. The results were analysed with my mentor, and based on this reflection together, we created the initial list (STARD_DI) to be sent to radiological experts for the first round of the Delphi technique. It was confirmed at this time that a second critical friend had agreed to participate in my study. This physician lived abroad in the UK. He was a radiologist who was well experienced in research methods and familiar with reporting tools such as the STARD.

In discussion with my mentor, we decided against sending the results from the needs assessment to both critical friends since there were not a lot of additional comments obtained from the participants. My mentor postulated that the reason for this might have been due to the fact that the radiological experts don't use STARD much and don't assess the quality of reporting diagnostic accuracy studies in this manner. Nonetheless, they were keen to participate and together we created the first round of the Delphi to be sent to the participants for their opinion.

4.7 Delphi Technique

Although the development of systematic tools such as STARD, CONSORT and QUADAS were developed by consensus with an expert panel utilizing Delphi methodology, the literature does not recommend or debate how this technique should be piloted (Clibbens et al., 2012). The advantage of utilizing this method is the latitude it affords the researcher, however; each iterative round of the Delphi must be carefully performed as poor utilization of the Delphi technique will weaken the overall scientific transparency and methodological rigor of the study (Hsu and Sandford, 2007).

As described by Hsu and Sandford (2007), data generated from the Delphi method can consist of both quantitative and qualitative data. Qualitative data is generated from open-ended questions, which are generally used with the first iteration of the Delphi; whereas subsequent iterations are conducted to collect enough data with the goal of reaching a consensus amongst the experts as the tool is built. Mixed methods and flexibility this tool offered was the reason why I used this data collection tool to conduct my action research project. With each round of questioning, the radiological experts were expected to respond individually, whereby their responses were later combined in an effort to meet consensus and the eventual construction of new knowledge. Quantitative analysis was done from each round to decide which items would remain in the next round. As the development of this tool was conducted in successive iterative Delphi rounds, the radiological experts were constantly reflecting on their existing knowledge to create the new tool. This new tool (RadSTARD) resulted in the generation of new knowledge that was accomplished via participatory action research and critical theory paradigm (Bunniss and Kelly, 2010).

4.7.1 Delphi Round 1

Items were added to the list such as the addition of a mixed standard or panel standard. The addition of this item came from my own independent readings and was approved by my mentor. Such an item may be pertinent in diagnostic accuracy studies where the reference standard is not specific (Bertens et al., 2013). The participants were also asked to rate whether or not sample size should be included in diagnostic accuracy results. On reflection, I was quite concerned about how many choices the radiological experts should have when providing their responses. In consultation with our methods centre statistician, and both critical friends, it was agreed that providing a Likert rating scale of 1-10 (1 least agree to 10 strongly agree) would render more statistical data for analysing later.

The first-round of the Delphi was sent to the radiologists via Survey Monkey. These questionnaires consisted of questions that would render both quantitative and qualitative responses, as they were open text boxes for the participant to provide additional comments. The results from this first round can be found in Chapter 5 on Project Findings.

4.8 Critical Friend #2

The second critical friend joined the study when the needs assessment and first round of the Delphi had been completed. He wondered how I came up with the questions for the needs assessment, I explained that the questions were created by me and sent to my mentor who had suggested that I send them to another physician, in this case my other "critical friend". He approved them and I sent the needs assessment out for response. Essentially, many of the items were accepted as being required by radiological experts, which upon reflection speaks to the fact that either the questions could have been worded differently, or some of the results rendered plasticity (Plous, 1993). As described by Plous (1993) plasticity refers to how one's attitudes and opinions are plastic; and therefore impact how they answer a question.

My mentor and I developed the revised STARD_DI after the needs assessment was done. This changed the format of the items for the STARD_DI; whereby instead of listing them with a corresponding page number like they are in the STARD, each item was posed as an open ended question, as recommended when conducting the Delphi (Hsu and Sandford, 2007)**.** There was also room for comments for some items. The initial draft version of STARD-DI provided the participant a frame of reference tailored to potentially reduce their concerns regarding the time required to participate in the project.

4.9 Delphi Round 2 and RadSTARD (Radiology Standards for Reporting of Diagnostic Accuracy)

Results from both rounds were reviewed with both critical friends. After the second round of the Delphi, consensus for which items to be maintained in the newly revised tool applicable to radiology was met by a cohort of radiological experts. The name of this newly revised tool developed by radiologists for radiologists was called RadSTARD (Radiology Standards for Reporting of Diagnostic Accuracy). Based on this list, an explanation and elaboration document was written for submission to REB for approval to validate the tool with the residents and Fellows.

4.10 Facing PAR Positionalities as an Insider-Outsider

After the needs assessment was completed, and the results reviewed with my mentor and critical friend, the first round of the STARD_DI Delphi was sent to the participants (Appendix 10). Within the email sent to them I requested that they complete the survey within two weeks. The survey link was sent to them via Survey Monkey, which ensures the anonymity of respondents. Throughout the two-week period, 'gentle reminder' emails were sent thanking those who had completed the survey and requesting that those who hadn't to kindly complete it at their next available opportunity. In the second follow-up email, I explained to the participants what they could expect after each round of responses had been analyzed. I was referring to the fact that with each Delphi iteration, they would be provided with a summary of the responses from the previous round in case they wanted to make any changes to their previous responses. This served two purposes; first, it prompted further reflection on their initial responses and second, with each round it eventually lead to meeting consensus on the items deemed required towards the development of the new tool. I would like to point out that the main reason I explained this to the participants is that I was considering the power differentials that occurs during PAR studies involving the Delphi technique (Fletcher and Marchildon, 2014). As collaborators, they were providing me with their expert opinion and knowledge, which was paramount to the development of the revised tool. Therefore, I wanted to as much as possible create dialogue with the collaborators as knowledge was being generated through interactions between me, the researcher, and those who agreed to participate (Fletcher and Marchildon, 2014).

4.11 Summary

In summary, after the needs assessment was completed, it took two rounds of the Delphi technique to finalize the development of the new tool, which was called the RadSTARD. An elaboration document was developed which described each item of the RadSTARD tool with reference to the STARD for those items that corresponded. Concomitantly, rationale for each item was provided that illustrated why reporting the item was recommended for radiology diagnostic accuracy trials.

Validation of this tool was conducted with the radiology residents and Fellows whereby; interim results were presented to them (Appendix 16). The next chapter further describes the steps involved in the development of RadSTARD and the results found upon validation of the tool.

5. Chapter 5: Project Findings

5.1 Study Conduct

The development of a new reporting tool specific to radiology diagnostic accuracy trials was conducted in collaboration with a group of physicians, and in two distinct phases. Firstly, a needs assessment was sent to eight radiological experts who had agreed to participate in the study. Based on their response, and feedback from my mentor and critical friends, the next phase of tool development was done via the Delphi technique. Each phase of the Delphi is akin to the cycles of action research; with each successive round, the panel of radiology experts invited to participate in the research has to critically reflect on their knowledge and experience when responding to the survey questions or when providing responses to qualitative questions (Hsu and Sandford, 2007, Koshy et al., 2011).

5.2 Phase One - Needs Assessment

Within critical literature, needs assessment is described as an informal method; however, physicians utilize a similar method to determine their learning needs with respect to their practice (Grant, 2002). Needs assessment surveys are invaluable in medicine as they are integral to continuous professional growth and development. Although as a stand-alone tool they do not constitute research, physicians are accustomed to using formal and informal methods to conduct needs assessments. For example, structured interviews and questionnaires have been used as needs assessment tools “for evaluation, assessment, management, education, and now appraisal and revalidation” (Grant, 2002:157).
Medical education can include various methods of undertaking a needs analysis: gap or discrepancy analysis, reflection on action, reflection in action, self-assessment by diaries, peer review, observation, critical incident review, significant event auditing, and practice review (Grant, 2002). Of the different types of needs assessment, I found reflection on action, and reflection in action to be the most relevant to my study, as the physicians need to reflect on their practice and/or knowledge when participating in the different phases of the study. As the radiological experts responded to the questions on needs assessment and the subsequent rounds of the Delphi technique, it required they reflect on their actions since their experiential learning would influence their responses (Grant, 2002).

A Likert Scale (Whiting et al., 2003) was provided whereby the radiological expert rated each item listed in the needs assessment according to the rating scale. Their options were to strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree. The following questions from the needs assessment were agreed upon if the combined responses for agree and strongly agree were > 70%. The survey responses for the needs assessment can be found in Appendix (25).

5.2.1 Needs Assessment Results

Responders: N = 8 Questions skipped: N = 0

Question	Response (>70%)
ELIGIBILITY	
Should the inclusion/exclusion criteria for enrolling the participants be included?	100%
PARTICIPANT SAMPLING	
Does knowing whether the study population was tested consecutively relevant to your interpretation of diagnostic accuracy studies?	75%
REFERENCE STANDARD	
The rationale for use of the reference standard should be stated	100%
Make and model of diagnostic imaging equipment should be provided	88%
How the test was delivered/performed should be provided	100%
METHODS	
Definitions for cut-off ranges should be provided	100%
Training and expertise of the person interpreting the index test and reference standard should be provided	100%
The interpreter should be blinded to the results of the previous tests	100%
The methods for calculating diagnostic accuracy between the index test and the reference standard should include confidence intervals to quantify uncertainty	100%
Should test reproducibility be provided?	75%
The time interval between the index test and the reference standard should be provided	100%
Should a description of those with the target disease be provided?	100%
Should other diagnoses or co morbidities be provided?	75%
Should a cross tabulation of the results for the index test and reference standard including indeterminate and missing results be provided?	88%
Should adverse events be reported?	100%
RESULTS	
Confidence intervals 95% are required to describe estimates of diagnostic accuracy and measures of statistical uncertainty	100%
Should indeterminate results be provided?	100%
Should estimates of variability between subgroups of participants be provided if done?	100%
Should estimates of test reproducibility be provided if done?	88%
The clinical applicability of the study findings should be provided	100%

(Bossuyt et al., 2003a)

5.2.2 Sensitivity and Specificity in the Title and Flow Diagrams

A Likert Scale was provided for the responders to rate their response as either strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree. These results indicate the percentage of responses not agreed upon by the cohort and that scored <70% when the agree and strongly agree responses were combined, the responders did not agree upon the following questions:

Question	Response (<70%)
Should diagnostic accuracy trials mention ‘sensitivity and specificity’ in the title of the study	13%
Is a flow diagram describing the number of participants who met eligibility criteria and went on to receive the index test and reference standard required?	63%

5.2.3 Recruitment of Study Participants

When radiological experts were asked which items they deemed most essential for recruitment of participants into a trial, it rendered the following response:

Item	1	2	3	Total	Average ranking
Presenting symptoms	25% 2	37.5% 3	37.5% 3	8	1.88
Results from previous tests	0% 0	37.5% 3	62.5% 5	8	1.38
Participants had received either the index test or the reference standard	75% 6	25% 2	0% 0	8	2.75

5.2.4 Items to Consider When Interpreting Radiology Diagnostic Accuracy Studies

Finally, the last needs assessment question resulted in qualitative data from four out of the eight respondents. Respondents were asked to indicate which additional items they thought should be included when interpreting diagnostic accuracy trials specific to radiology, with the following results:

#	Responses
1	What part of the result is more generalizable for other equipment? Flow chart for practical implementation and quality control
2	Precision of the test – inter-observer agreement. Comment on the sample size and practical limitations. If readers were blinded and if there was a training session
3	Differences in outcomes. This is the main objective of these types of studies. Was the time different? Could contrast have been avoided? Did the diagnosis change? Was less radiation used? Etc.
4	Number of patients who had the reference standard test after the index test. Inter-and intra-observer variability

5.3 Results of the Needs Assessment

The results of the needs assessment indicated that most items from the current STARD tool as relevant when reporting the quality of diagnostic accuracy studies, presenting a challenge.

At this stage of research, I consulted with critical friend #1 who offered suggestions on the Delphi technique phase of the study and how to group the items on the STARD- DI (Standards for Reporting of Diagnostic Accuracy - Diagnostic Imaging) checklist for radiological experts to offer their opinion on. It was pointed out that some of the items from the needs assessment may not truly be relevant to radiology, and suggested that I make the list more succinct to interpreting diagnostic accuracy. For example, giving an example of the study title and adding the words “sensitivity and specificity” to the titles to determine whether or not it is really needed. This is the first item of the STARD checklist. He also felt that the second question of the STARD checklist was redundant; do we really need to determine if the research question was stated? In addition, he also suggested that space be provided for the radiologist to add their own comments with an opportunity to rate each item as a yes or no response as opposed to the page numbers option in the current format of the STARD tool. This is more relevant to editors for publishing and pointed out that the purpose of the tool I was developing is for the clinician to use when interpreting diagnostic accuracy studies specific to radiology.

It wasn't until the end of the needs assessment survey that my mentor identified a second critical friend to collaborate with on my research. The results of my needs assessment were discussed with my mentor including my concerns that not a lot of additional comments were provided. The mentor felt this could be attributed to the fact that radiological experts don't used the STARD tool much and are therefore, not accustomed to assessing the quality of the reporting of the literature in this manner.

Nonetheless, they were keen to participate in the study and we moved onto the Delphi technique phase (Hsu and Sandford, 2007). This phase required that the respondents provide feedback with each cycle, encouraging critical reflection and resulting in the collection of additional data. Ultimately, this lead to the identification of which items should be maintained from the current STARD tool in the revised tool and delineated which items were different and specific to radiology diagnostic accuracy trials.

5.4 STARD_DI Delphi 1 Consensus Met

Similar to the needs assessment, the following questions were agreed upon if the combined responses for agree and strongly agree were > 70%. A Likert Scale was provided whereby the radiological expert rated each item according to the rating scale of strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree.

The survey responses for the STARD_DI Delphi Round 1 can be found in Appendix 27.

5.4.1 STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

Item		Response N = 8
Title & Abstract	Should Radiology Diagnostic Accuracy Trials include the words "Sensitivity and Specificity" or "Diagnostic Accuracy" in the Title/Abstract?	75%
	"Diagnostic Accuracy" in the Title/Abstract?	100%
	Should Radiology Diagnostic Accuracy Trials explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition?	100%
METHODS **Patient selection, recruitment and sampling**	Should Diagnostic Accuracy Trials provide inclusion and exclusion criteria for patients?	100%
	Details of setting and location of study (e.g. whether primary or secondary care)?	75%
	Whether data collection is prospective or retrospective?	100%
	Whether patient selection was consecutive or non-consecutive?	100%
	Sample size and limitations?	100%
	Start and end dates of the study?	100%
Diagnostic test methods	Should Radiology Diagnostic Accuracy Trials explicitly state the reference standard and its rationale?	100%
Imperfect reference standard	If the Reference Standard is unavailable or imperfect, should trials state what alternative was used to justify the alternative? For example: New MRI technique for rotator cuff tear (index test) compared to arthroscopy (reference standard). If only 10% went on to surgery this would create a verification bias. Test only works in patients with severe disease creating a verification bias. Therefore would a mixed standard of reference be required if several orthopedic surgeons think there is no rotator cuff injury?	100%
TEST	Technical specifications for the index and reference test should be reported in all radiology diagnostic accuracy studies	75%
	Sound theoretical physics basis of the index test should be provided for new techniques	88%
	If the test was modified during the study, do you need to know?	100%
Analysis	Should Radiology Diagnostic Accuracy Trials report cut-off values for specific diagnostic criteria for index and reference tests?	100%
	Training and number of investigators?	100%
	Whether extra training was provided for a new technique?	100%
	Whether readers were blinded to prior test results?	100%
	Whether readers were blinded to clinical information?	100%
RESULTS	Should Radiology Diagnostic Accuracy Trials report study flow, including eligible patients who did not undergo index or reference tests, and explained why?	100%
	Should a flow diagram be provided?	71%
	Should patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies be provided?	100%

	Should the severity (spectrum) of the disease entity be explicitly reported?	75%
	Should the time difference between the index test and reference test be provided?	100%
	Do you need to know if any other treatments were provided between the two tests?	87.5%
Adverse Events	Should adverse events be reported for either the index test or reference standard?	87.5%
STATISTICS	Radiology Diagnostic Accuracy Trials one should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI)	100%
	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers; and describe how this data was handled	87.5%
	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability	100%
DISCUSSION	Should the clinical relevance of the study findings be provided?	87.5%

(Bossuyt et al., 2003a)

5.4.2 STARD_DI Delphi 1 Items – Consensus Not Met

The same Likert Scale was provided for the first round of the STARD_DI Delphi: strongly disagree, disagree, neither agree, nor disagree, agree, strongly or agree. These results indicated the percentage of responses that were not in agreement by the cohort with a score of <70% for agree and strongly agree. The responders did not agree on the following questions:

5.4.3 STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

Item		**Response N = 8**
Title	"Sensitivity and Specificity" in the Title/Abstract?	50%
Methods	Whether a power calculation was performed?	38%
	Should a thorough description of the nature of the disease be presented?	38%
	Only where the test is not the standard of care?	38%
Test	Should Radiology Diagnostic Accuracy Trials explicitly state generalizability of the technique to other vendor equipment?	63%
	For all techniques used?	25%
Analysis	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including > 3 observers with variable expertise?	50%
Results	Should changes in diagnosis after each test be reported?	50%
	Should it be reported if contrast could have been avoided?	50%
	Would you need to know if the level of radiation was less for the index test?	50%
Statistics	Should Radiology Diagnostic Accuracy Trials include a cross tabulation of results of index and reference tests?	63%
	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility?	63%
	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?	25%
	Should this be provided all the time?	13%
	Technical specifications for the index test and reference tests should be reported only when the test is not standard of care?	38%

(Bossuyt et al., 2003a)

5.5 STARD_DI Delphi 1 Items in the Title and Abstract

When asked: “What should be included in the Title and Abstract?” the responses were as follows:

5.5.1 STARD_DI Delphi Round 1 Results

Responders: N = 4 answered and 4 skipped this question

1. Prevalence of the disease impact in management
2. At least accuracy (or equivalent term), and both test method(s) and used gold standard
3. Depends on the study, but “Diagnostic Accuracy” is usually appropriate
4. The purpose of the study

5.5.2 STARD_DI Additional Items Recommended by the Expert Panel

The following response rate was provided with respect to which additional items the panel indicated should be included in the revised checklist for Radiology Diagnostic Accuracy studies:

5.5.3 STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

Item		Response N = 8
Title	Does this need to be included in a checklist for Radiology Diagnostic Accuracy?	63%

STARD_DI Delphi Round 1 Results

Responders: N = 1 answered and 7 skipped this question

5.5.4 STARD_DI Delphi 1- Imperfect Reference Standard

Only one responder answered the following question:
If you do not think a mixed standard or panel standard would be of benefit, please insert your comments of what you think would be a viable alternative for diagnostic accuracy studies with an imperfect reference standard?

1. There is almost zero ways to have a perfect reference standard, and radiological studies will mostly be biased by selection. Still the accuracy applies to the selected population under that diagnostic question.

The following responses are a continuation of the items from the first round of the STARD_DI Delphi and were not agreed upon. The first question is in response to the qualitative responses listed above. Although only one participant provided a detailed response with respect to an imperfect reference standard, 75% of the respondents agreed that the technical specifications for the index test and reference standard should be provided in the results of radiology diagnostic accuracy trials. In addition, in case they wanted to change their response based on their cohort's responses, the respondents were also provided a summary of the responses from round 1 of the Delphi, which were anonymized.

5.6 STARD_DI Round 2

Once the responses were received, they were analyzed in conjunction with my mentor and critical friends. For the next round of the Delphi, we removed the responses for items that scored 100% from the panel. However, where the responses were at 70% panel agreement, a positive consensus was met as described by (Nieuwenhuijze et al., 2014), whereas a negative consensus was described as < 70% panel agreement.

Hence, items that scored < 70% were added to the next round and the respondents were asked to rank the items they deemed to be a priority. This rendered patterns of agreement and disagreement as consensus of the revised tool developed (Hsu and Sandford, 2007). In addition, the respondents were also provided a summary of their responses from round 1 of the Delphi in case they wished to change their responses based on their cohort's responses (anonymized).

One critical friend agreed with this method, whereas to ensure the panel understood the question, the second critical friend suggested presenting supporting evidence for inclusion of each contentious item in Round 2. I explained that items sent back to the panel were based on the prior tool (STARD) (Bossuyt et al., 2003b) and at this point of development, not based on supporting evidence per individual item.

The critical friend also suggested, in addition to the median and distribution, listing the personal score given by the panel participant (anonymized to other panel members). The panel member could then recall how he/she scored the item and consider whether to revise their score towards consensus or retain their original score.

This could not be executed and represents a limitation of the survey tool, which will be discussed further later. The benefit of the Delphi method is that radiological experts were reflecting on their professional practice and tacit knowledge when providing their responses during the development of the new tool (Eraut, 2000).

5.6.1 STARD_DI Delphi Round 2

With the second round of the Delphi, participants were provided a summary of their responses from the round and offered an opportunity to change any of their responses based on their cohort's response. One of the participants decided to revise their response based on the first round of the Delphi.

5.6.2 STARD Delphi 1 Revised Reponses – Consensus Met

The number of items that scored greater than 70% positive rate did not change from the first round of the Delphi 1; however, the preponderance of positivity between agree or strongly agree changed for a few items as listed below. There were 9 responders in the revised responses from the Delphi Round 1. It should be clarified that, when accessing the same survey, only one member of the expert panel chose to change their response resulting in a total of nine responders. We were not able to identify the change or duplication due to the anonymity of the survey.

STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

STARD_DI Delphi Round 1 Revised Reponses

Responders: N = 9 Questions skipped: N=1

Item		Response N = 8	Response N = 9
Title & Abstract	Should Radiology Diagnostic Accuracy Trials include the words "Sensitivity and Specificity" or "Diagnostic Accuracy" in the Title/Abstract?	75%	78%
	"Diagnostic Accuracy" in the Title/Abstract?	100%	89%
	Should Radiology Diagnostic Accuracy Trials explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition?	100%	100%
METHODS Patient Selection, recruitment & sampling	Should Diagnostic Accuracy Trials provide inclusion and exclusion criteria are patients?	100%	100%
	Details of setting and location of study (e.g. whether primary or secondary care)?	75%	78%
	Whether data collection is prospective or retrospective?	100%	100%

	Whether patient selection was consecutive or non-consecutive?	100%	100%
	Sample size and limitations?	100%	100%
	Start and end dates of the study?	100%	100%
Diagnostic test methods	Should Radiology Diagnostic Accuracy Trials explicitly state the reference standard and its rationale?	100%	100%
Imperfect Reference Standard	If the Reference Standard is unavailable or imperfect, should trials state what alternative was used to justify the alternative? For example: New MRI technique for rotator cuff tear (index test) compared to arthroscopy (reference standard). If only 10% went on to surgery this would create a verification bias. Test only works in patients with severe disease creating a verification bias. Therefore would a mixed standard of reference be required if several orthopedic surgeons think there is no rotator cuff injury?	100%	100%
Test	Technical specifications for the index and reference test should be reported in all radiology diagnostic accuracy studies	75%	78%
	Sound theoretical physics basis of the index test should be provided for new techniques	88%	78%
	If the test was modified during the study, do you need to know?	100%	100%
Analysis	Should Radiology Diagnostic Accuracy Trials report cut-off values for specific diagnostic criteria for index and reference tests?	100%	89%
	Training and number of investigators?	100%	100%
	Whether extra training was provided for a new technique?	100%	100%
	Whether readers were blinded to prior test results?	100%	100%
	Whether readers were blinded to clinical information?	100%	100%

(Bossuyt et al., 2003a)

STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

STARD_DI Delphi Round 1 Revised Reponses

Responders: N = 9 Questions skipped: N = 1

Item		**Response N = 8**	**Response N = 9**
RESULTS	Should Radiology Diagnostic Accuracy Trials report study flow, including eligible patients who did not undergo index or reference tests, and explained why?	100%	100%
	Should a flow diagram be provided?	71%	75%

	Should patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies be provided?	100%	100%
	Should the severity (spectrum) of the disease entity be explicitly reported?	75%	78%
	Should the time difference between the index test and reference test be provided?	100%	100%
	Do you need to know if any other treatments were provided between the two tests?	87.5%	89%
Adverse Events	Should adverse events be reported for either the index test or reference standard?	87.5%	89%
Statistics	Radiology Diagnostic Accuracy Trials one should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI).	100%	100%
	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers; and describe how this data was handled	87.5%	89%
	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability	100%	100%
DISCUSSION	Should the clinical relevance of the study findings be provided?	87.5%	89%

(Bossuyt et al., 2003a)

5.6.3 STARD Delphi 1 Revised Reponses – Consensus Not Met

The same Likert Scale was provided: strongly disagree, disagree, neither agree, nor disagree, agree, strongly or agree. These results indicate the percentage of responses that were not in agreement by the cohort with a score of <70% for agree and strongly agree. The responders from the first round of the STARD_DI Delphi did not agree upon the following questions:

STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

STARD_DI Delphi Round 1 Revised Reponses

Responders: N = 9 Questions skipped: N = 1

Item		**Response N = 8**	**Response N = 9**
Title & Abstract	"Sensitivity and Specificity" in the Title/Abstract?	50%	56%
Patient Selection, recruitment & sampling	Whether a power calculation was performed?	38%	44%
	Should a thorough description of the nature of the disease be presented?	38%	44%
	Only when the test is not the standard of care?	38%	33%

	Should Radiology Diagnostic Accuracy Trials explicitly state generalizability of the technique to other vendor equipment?	63%	67%
	For all techniques used?	25%	33%
Analysis	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including > 3 observers with variable expertise?	50%	44%
RESULTS	Should changes in diagnosis after each test be reported?	50%	56%
	Should it be reported if contrast could have been avoided?	50%	56%
	Would you need to know if the level of radiation was less for the index test?	50%	67%
	Should Radiology Diagnostic Accuracy Trials include a cross tabulation of results of index and reference tests?	63%	67%
STATISTICS	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility	63%	67%
	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?	25%	22%
	Should this be provided all the time?	13%	11%
	Technical specifications for the index test and reference tests should be reported only when the test is not standard of care?	38%	33%

(Bossuyt et al., 2003a)

5.6.4 STARD_DI Additional Items in the Title and Abstract

When asked: “What should be included in the Title and Abstract?” the one additional response was:

1. Clinical impact, adaptability to different population

5.6.5 STARD_DI Additional Items Recommended by the Expert Panel

The following response rate is in response to the qualitative response listed above, with the results from the STARD_DI Delphi 1 compared to the revised response when the survey was completed again:

STARD_DI Delphi Round 1 Results

Responders: N = 8 Questions skipped: N = 0

STARD_DI Delphi Round 1 Revised Reponses

Responders: N = 9 Questions skipped: N = 1

Item		**Response N = 8**	**Response N = 9**
Title	Does this need to be included in a checklist for Radiology Diagnostic Accuracy?	63%	44%

5.7 STARD_DI Delphi Round 2

Based on their responses from Delphi Round 1, there were seven questions in the second round of the Delphi whereby participants were asked to choose three out of the seven items listed they think should remain in the revised checklist for STARD_DI. To encourage the respondents to selectively choose which items they thought were more relevant, it was my mentor who suggested the respondents choose 3 out of 7 items.

In addition, respondents were also provided the opportunity to add comments stating their rationale concerning the items they rated as a priority amongst the items reviewed.

Only the items that did not reach consensus were provided to the participants via the Survey Monkey method. The same Likert Scale as before was provided and radiological experts rated each item according to the rating scale of strongly agree, disagree, neither agree, nor disagree, agree, strongly or agree. The survey responses for the STARD_DI Delphi Round 2 Revised can be found in Appendix 28. The following two items were agreed upon as the combined responses for agree and strongly agree were > 70%:

5.7.1 STARD_DI Delphi Round 2 Consensus Met

Responders: N = 8 Questions skipped: N = 0

Item		Response N = 8
	Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise	75%
Rationale	1. Inter-observer agreement is valuable, but as long as the level of expertise is explicitly stated. It should not be mandatory to provide more than two observers 2. Ideal but sometimes not practical	
	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility	83.3%
Rationale	1. None provided	

5.7.2 STARD_DI Delphi Round 2 Consensus Not Met

Responders: N = 8 Questions skipped: N =0

The remaining five items did not reach consensus, as the response rate was <70%.

STARD_DI Delphi Round 2 Results:

Answered: N = 6 Questions skipped: N = 2

Item		Response N = 6
	Whether a power calculation was performed	50%
Comment	1. Need to know how many patients required to have statistical significance 2. Not always possible to provide a power calculation, especially for new diagnostic techniques or for populations who have not been evaluated previously 3. Statistical game sometimes	
	Radiology Diagnostic Accuracy Trials should explicitly state generalizability of the technique to other vendor equipment for all techniques used.	20%
Comment	1. A study shouldn't be done unless the test can be generalized to all vendor platforms 2. Readers should be alerted if the performance of the technique is vendor specific or limited to a specific population 3. May be difficult to guess sometimes	37.5%
	Changes in the diagnosis after each test should be reported.	60%
Comment	1. There were no responses	
	You should know if the level of radiation was less for the index test.	33.3%
Comment	1. Should be reported if test involves radiation 2. Usually not a game changer	
	Radiology Diagnostic Accuracy Trials should explicitly state generalizability of the technique to other vendor equipment for all techniques used.	66.7%
Comment	1. I find this question confusing	

5.8 The RadSTARD (Radiology Standards for Reporting Diagnostic Accuracy) Tool

The results from both rounds of the Delphi technique where consensus for items scoring >70% were combined in a new checklist, and the revised STARD_DI tool was renamed the RadSTARD tool. Although the results gathered resulted in many of the items from the original STARD tool being maintained, several items specific to the reporting of radiology diagnostic accuracy studies were added to this revised tool and the name changed to RadSTARD.

5.9 Differentiating the STARD Checklist from the RadSTARD Checklist

The final checklist/tool was 28 items. I combined item 5 (data collection) to include the start and end-dates, as it did not seem necessary to separate the two. The collection of study data with the start and end-date of the research initiative is routinely described in the literature and supplying the start and end-date of any study is an ethical requirement for any submission to the local research ethics board. Based on the two rounds of the Delphi, there were 13 items from STARD_DI that did not meet consensus <70% agreement. Notably, 2 of the items from the current STARD were not in the RadSTARD: methods for calculating test reproducibility and the provision of a cross tabulation of the results of the index tests and reference standard.

The items most specific to the reporting and interpreting of radiology diagnostic accuracy studies were the following:

1. Including the word 'diagnostic accuracy' in the title of this study
2. Explicitly stating that the aim of this study is to compare the index test with the reference standard to diagnose a specific condition
3. With respect to eligibility criterion, the provision of information as to whether or not the radiology diagnostic accuracy study relates to primary or secondary care should be specified
4. Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard. This item differs from the original STARD checklist as the panel were asked to rank order the importance of this item. Therefore, information pertaining to presenting symptoms and results from previous tests were ranked lower in priority and therefore not included in the revised tool
5. Data collection and the start and end dates were combined as one item
6. The provision of sample size and limitations when reporting the results of radiology diagnostic accuracy studies
7. If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified
8. Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies
9. Sound theoretical physics basis of the index test should be provided for new techniques
10. Radiology Diagnostic Accuracy Trials should report cut-off values for specific diagnostic criteria for index and reference tests
11. Modifications during the conduct of the radiology diagnostic accuracy study should be reported if they occurred
12. The training and number of investigators should be described including any extra training for new techniques. Emphasis on new techniques

5.10 RadSTARD –Radiology Standards for Reporting of Diagnostic Accuracy Tool

Item	
1. Title & Abstract	**Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract**
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition**
3. METHODS Patient selection, recruitment & sampling	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard**
5.	**Data collection (prospective or retrospective) should be provided with start dates and end dates provided**
6.	Whether patient selection was consecutive or non-consecutive?
7.	**Sample size and limitations should be provided**
8. Diagnostic test methods	Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale
9. Imperfect reference standard	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified**

10. **Test**	**Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies**
11.	**Sound theoretical physics basis of the index test should be provided for new techniques**
12.	**Modifications during the study should be reported if they occurred**
13. **Analysis**	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests
14.	The training and number of investigators should be described including **any extra training for new techniques**
15.	Whether readers were blinded to prior test results or clinical information should be known
16. **RESULTS**	Radiology Diagnostic Accuracy Trials report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations
17.	Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies should be provided
18.	The severity (spectrum) of the disease entity should be explicitly reported
19.	The time difference between the index test and reference test should be provided
20.	Details of any other treatments provided between the two tests should be described
21. **Adverse Events**	Adverse events should be reported for either the index test or reference standard
22. **STATISTICS**	Radiology Diagnostic Accuracy Trials one should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI)
23.	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers and describe how this data was handled.
24.	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability

25.	Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise
26.	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility
27. **DISCUSSION**	The clinical relevance of the study findings should be provided

Items in bold (1, 2, 3 (partial), 4, 5, 7, 9, 10, 11, 12 and 13 (partial) and 14 (partial) are specific to the RadSTARD Reporting Tool.

Bossuyt (2003)

5.11 Current STARD Compared to the RadSTARD

STARD checklist for reporting of studies of diagnostic accuracy

(January 2003 version)

Section and Topic	**Item #**		**Compared to RadSTARD**
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity')	Different
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	Different
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected	Partially different
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	Different
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	√
	6	Data collection: Was data collection planned before the index test and reference standard performed (prospective study) or after (retrospective study)?	Different
Test methods	7	The reference standard and its rationale	√
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard	Partially different

	9	Definition of, and rationale for, the units, cut-offs and/or categories of the results of the index tests and the reference standard	Partially different
Section and Topic	**Item #**		
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard	Different
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers	√
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals)	√
	13	Methods for calculating test reproducibility, if done	Not included
RESULTS			
Participants	14	When the study was performed, including beginning and end-dates of recruitment	Combined with item 6
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms)	√
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended)	√
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between	√
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition	√
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard	Not included
	20	Any adverse events from performing the index tests or the reference standard	√
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)	√
	22	How indeterminate results, missing data and outliers of the index tests were handled	√
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done	√
	24	Estimates of test reproducibility, if done	√
DISCUSSION	25	Discuss the clinical applicability of the study findings	√

(Bossuyt et al., 2003a)

Legend: √ - RadSTARD items similar to STARD
Remaining items are as described when compared to the STARD – different, partially different or not included in the RadSTARD

5.12 Validation of the RadSTARD

Once the RadSTARD tool was developed, it was validated in two phases. Conducting validation in two phases, or cycles, was done so that those participating (radiology residents and Fellows) could reflect on their responses from the initial phase of assessing RadSTARD. The following results summarize the findings from these two phases:

5.12.1 Phase 1 – Interpretation of Literature with RadSTARD Checklist (Appendix 17)

After the radiology residents and Fellows read the article entitled, “Comparison of diagnostic accuracy of magnetic resonance imaging and multidetector computed tomography in the detection of pelvic fractures” by Henes et al. (2012), they provided their level of confidence when interpreting this article in reference to the RadSTARD tool and elaboration document (Appendix 4).

RadSTARD consists of 25 items and those who agreed to participate rated their level of confidence via a Likert Rating Scale, which ranged from 1-10 (where 1 is least likely and 10 is most likely). For this phase, 18 (58%) radiology residents and 15 (88%) Fellows participated.

Once the raw data was collected, data analysis was performed by a statistician from the Methods Center at The Ottawa Hospital. The goal was to test the variable data based on the survey responses to determine if there is a central tendancy between the radiology residents and Fellows. A non-parametric test called the Mann-Whitney Test ranking individual scores of a given sample and then comparing the results between the two groups was used to accomplish this (Dawson, 2012). The Mann-Whitney analysis was completed via the NPAR1WAY procedure used to conduct a standard analysis of variance on the raw data based on the distribution of results between the two groups (Institute, 2012).

5.13 RadSTARD Quantitative Analysis: Mann-Whitney Results

The general idea of the test was to assume that the distributions of values for radiology residents and Fellows were similar. If this was truly the case then values in one group should not rank systematically higher than values in the other. If they did, this would have rendered evidence that one of the groups had shifted relative to the other or in different locations of the distributions. The mean score was the mean rank from each group.

The cut-off for interpretation on the Likert Scale was seven. Radiology residents tended to provide scores of seven or higher more often in comparison to the Fellows. The Mann-Whitney (or Wilcoxon) test is a non-parametric analog to the t-test for situations were making assumptions about normality or parameter values (mean, SD) may be problematic.

The entire data set was ranked from smallest to largest value. A test statistic was used to compare the sum of the ranks between two groups. The general idea of the test was to assume similar distributions of values for radiology residents and Fellows. If this was truly the case, then the values in one group should not rank systematically higher than values in the other. If they did, this would be evidence that one of the groups shifted relative to the other, rendering different locations of the distributions. This is one non-parametric analog to the t-test where the means represent the location of each group's distribution and the idea is to parametrically test for a difference in location. RadSTARD item #18 was the closet p-value of significance between the two groups. The results from the Mann-Whitney can be found in Appendix 29.

5.14 RadSTARD Quantitative Analysis: Chi-Square and Fisher Exact Tests

Chi-Square (or Fisher exact tests if needed) was completed to assess the proportions expressing confidence for each question from the raw data. The key results are descriptive cross-tabulations of the responses for each question, with a p-value indicating whether or not there is a difference in the proportions expressing confidence for that question between the fellows and residents (differences were significant for two of the items).

The p-value between the two groups was not significant. Whenever any warning appeared from the Chi-Square test, the p-value from the Fisher exact test was used to look for an association between the response categories between the two groups (radiology residents/Fellows). RadSTARD Question 7 and Question 9 resulted in the only differences for the chi-square. The results from this analysis can be found in Appendix 30.

In summary, the chi-square analysis resulted in statistical difference between the two proportions for items # 7 and 9 of the RadSTARD. Item # 7 of RadSTARD (sample size) and item # 9 (knowing if an alternate reference standard was described) ranked higher in confidence level in interpreting the article by the Fellows as compared to the radiology residents. This result could be explained simply due to their level of training and exposure to research and interpretation of published literature since Fellows are at a more advanced level of training than radiology residents. The results of the Mann-Whitney resulted in no statistical significance between the two physician groups except for item # 18 of RadSTARD pertaining to knowledge of the severity of a disease state that renders a p-value of 0.0915. As there was no evidence of any significant difference between the two groups, we can conclude the RadSTARD rendering generalizability as a tool that can be used to interpret the results of radiology diagnostic accuracy studies.

5.15 Phase 2 – Measuring the Usefulness of the RadSTARD Tool

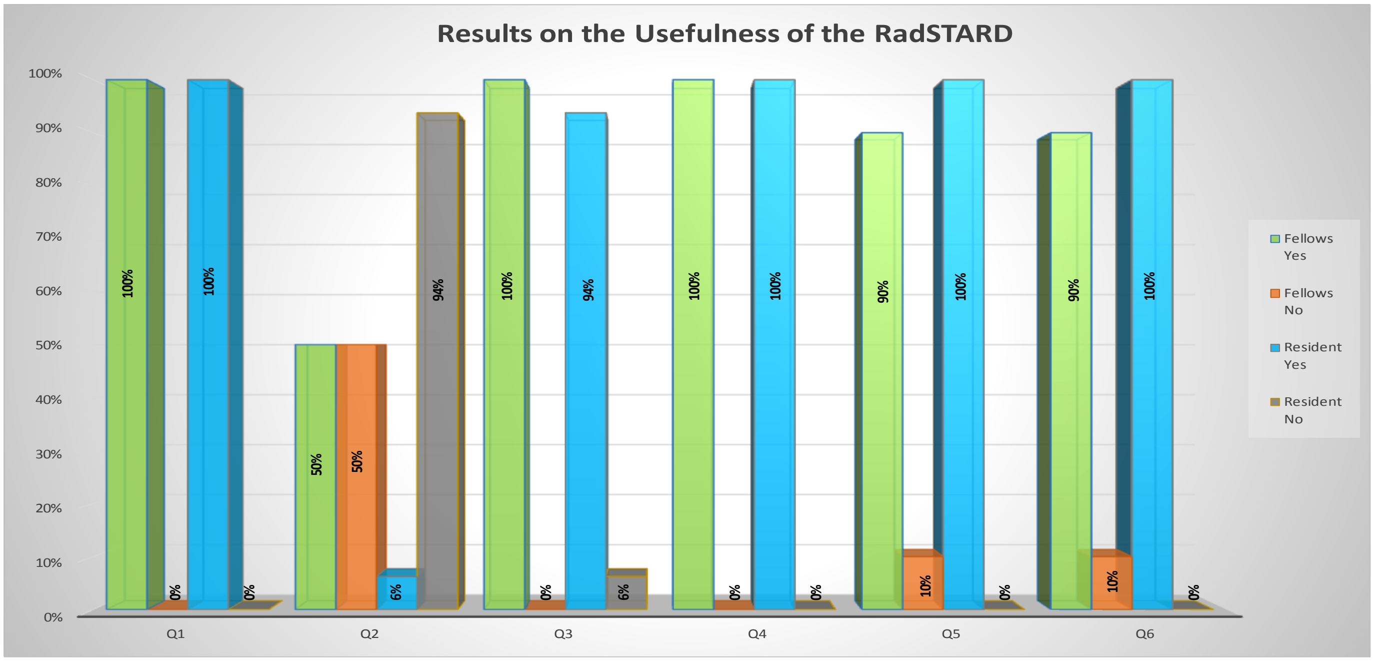
This phase or cycle of assessment was completed after the raw data from the results was presented to the radiology residents and Fellows via a power-point presentation. See Appendix 16. The purpose of this assessment was to obtain qualitative responses from radiology residents and Fellows on the usefulness of RadSTARD when interpreting radiology diagnostic accuracy literature. Those that agreed to participate in this follow-up cycle from Phase 1 provided their responses to six questions on the usefulness of the RadSTARD tool plus elaboration document. For this phase, responses were obtained from 16 radiology residents and 10 Fellows.

5.16 Table 3 - The Usefulness of the RadSTARD Questionnaire

Q1: Did the RadSTARD cover all the important items?

Q2: Were any RadSTARD items omitted, added, or modified?

Q3: Was the elaboration document easy to understand?



Q4: Did you find RadSTARD useful in rating the quality of the reporting for the diagnostic accuracy article?

Q5: Did the provision of the RadSTARD tool increase your level of confidence to interpret the article?

Q6: Would you use RadSTARD again?

5.16.1 The Usefulness of the RADSTARD Questionnaire (Appendix 19)

This follow-up questionnaire was provided to radiology residents and Fellows after using the RadSTARD tool, elaboration document and 2-page summary RadSTARD document to interpret the radiology diagnostic accuracy study by (Henes et al., 2012).
Another diagnostic accuracy by McComiskey et al. (2012) that was reported as per STARD recommendations was provided as reference, along with the STARD checklist and elaboration document for comparison.

This phase of validation for RadSTARD was performed in an effort to obtain qualitative data as space was provided for the respondents to enter their additional comments. The overall response rate for the residents and Fellows in completing this phase of validation for RadSTARD was 51% for radiology residents, and 58% for Fellows.

The Usefulness of the RadSTARD Questionnaire consisted of 6 questions and respondents were asked to critically reflect on the RadSTARD tool they used it in the first cycle to rate their level of confidence in interpreting the article by (Henes et al., 2012).

The questions and response rates for both groups of physicians were as follows:

5.16.2 Table 4 - Usefulness of the RadSTARD Questionnaire – Raw Data

Question	**Radiology Residents** **N = 16**	**Radiology Fellows** **N = 10**	**Comments/observations**
Q1: Did RadSTARD cover all the important items?	100%	100%	No additional comments provided
Q2: Were any RadSTARD items omitted, added or modified?	93%	50%	Only one of the Fellows provided additional comments here, which will be discussed below

Q3: Was the elaboration document easy to understand?	93%	100%	Two residents provided dichotomous comments on this document
Q4: Did you find RadSTARD useful in rating the quality of the reporting for the diagnostic accuracy article?	100%	100%	No additional comments provided
Q5: Did the provision of the RadSTARD tool increase your level of confidence to interpret the article?	100%	90%	No additional comments provided
Q6: Would you use the RadSTARD again?	100%	90%	No additional comments provided

5.17 Additional Qualitative Data – Radiology Residents

The Usefulness of the RadSTARD Questionnaire provided space for the radiology residents and Fellows to insert additional comments, encouraging feedback.

5.17.1 Radiology Residents Comments

As per the Usefulness of RadSTARD
Questionnaire Q3: Was the elaboration document easy to understand?
Resident 1 - One radiology resident asked to "please shorten" the length of the elaboration document.
Resident 2 - Another radiology resident commented, "the summary document provided an excellent guide."

Field notes were recorded throughout the entire study. After presenting the interim analysis of raw data from Phase 1 – the results were presented to both physician groups. At this phase, notes from my journal described one resident expressing how he thought all items of RadSTARD were very relevant to interpreting the article with confidence. He was surprised that some responses were low. In this example he admitted to knowing the sample size of the study population in the article they read. Knowing this information as they interpret the article with RadSTARD resulted in a 38% confidence level rate for the radiology residents whereas the Fellows rated this item significantly higher with a 92.3% confidence level.

As I reflected-in-action to his response, I explained to the radiology resident that the RadSTARD tool was developed in consensus with radiological experts who decided via the Delphi technique which items met consensus versus which items did not. Adding sample size to the RadSTARD tool was an item, not included in the original STARD tool. He was surprised that the radiology residents ranked it so low compared to the Fellows. However, I myself attributed this to their level of training and exposure to research.

As per the RadSTARD elaboration document, when conducting a diagnostic accuracy trial the results of the index test are compared to the results from the reference or gold standard test that each participant undergoes. Therefore, to render the results statistically significant, it is pertinent to know how many participants need to undergo both tests. Knowing the sample size and its limitations will provide the confirmatory information that the results from the index test and the reference standard have statistical power (Riegelman, 2013).

Given that confidence intervals are routinely provided in the results of clinical trials and diagnostic accuracy studies, this value can result in a larger sample size (Riegelman, 2013). If the sample size has a large positive population, this will render increased precision around the confidence interval for sensitivity (Strassle et al., 2012). Determining whether the results of diagnostic accuracy studies have statistical power is very different from hypothesis testing investigations where there is no hypothesis that the index test is better than the reference standard. It is appropriate to determine what measurements around the confidence interval for sensitivity and specificity would be clinically relevant. This is routinely done for hypothesis testing when tests are compared to each other. In order to render a substantial significant power between the two tests requires a large sample size. Sample sizes are calculated for case studies and may be used as a guide for diagnostic accuracies studies. For example, sample sizes of 100 to several hundred participants with or without the disease entity under study is an appropriate sample size for evaluating diagnostic tests. This is especially true when the pre-test probability of the disease entity among those who undergo the tests is greater than 50%. Alternatively, the sample size for a screening test would require a much larger sample size such as those found in cohort studies or randomized controlled studies (Riegelman, 2013).

If a diagnostic accuracy study was being done to determine if a newer imaging technique was better than a conventional (reference standard) test, it would be of interest to know if the difference between the two tests was statistically significant. If the difference between the two tests was not deemed to be significantly different, this would clarify with certainty that there was no clinically important difference, rather than postulating that there may have been an important difference to note but not possible to tell due to the lack of an appropriate sample size (Eng, 2004).

Sample-size calculation should be performed when the protocol is being developed. In general, research studies that are properly designed include how statistical power was defined for the particular medical condition or disease entity under study. This includes calculating a sample size. Conversely, in most medical imaging journals sample size calculation is not provided. The provision of the α error (0.05) should be provided in the section on statistical analysis, implying the probability of false positive results. Therefore, if the results of the study render no statistical significance between the two groups, not supplying a sample size and power analysis becomes an issue (Sardanelli and Di Leo, 2009a).

Appropriate Sample Size Parameters

An appropriate sample size is comprised of the following five study parameters:

1. Minimum expected difference which is also referred to as the effect size
2. Estimated measurement variability which is the standard deviation that is expected between measurements made amongst the interpreters
3. Statistical power - as the desired study power increases the sample size also increases
4. Significance criterion which is typically set to 0.05
5. Statistical analysis - one versus two tailed analysis (Eng, 2003)

Therefore, a sample size to be included with the study protocol should be calculated a priori, and the definition for the sample size derived by the minimal difference thought to render a clinical impact. This number is typically derived by radiologists from a critical analysis of the literature, before conducting the study (Sardanelli and Di Leo, 2009a).

5.18 Additional Qualitative Data – Radiology Fellows

After the Fellows were introduced to the RadSTARD tool, the feedback rendered was positive:

Fellow 1 - very excited to learn about this list as he had recently completed a course in Australia on how to interpret a paper. He added that he wanted to know “how can you decide after you read an article if you can follow the advice of the paper or not”.

Hence, he was very enthusiastic about reading the article and completing the questionnaire.

Fellow 2 – upon reading the article and completing the survey rating his confidence level in interpreting the article, he stated, “he found the RadSTARD was a road-map to interpreting the literature”.

Fellow 3 – the only Fellow that provided constructive feedback on several items from the RadSTARD tool as listed below:

RadSTARD Item 3: Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. whether primary or secondary care) should be provided

The Fellow commented that the "prevalence of disease in the study population needs to be provided. India has a high prevalence for tuberculosis. Japan has a high prevalence for gastric cancers. Different prevalence's would have an impact on the general acceptability and adoption of the said index test."

I appreciated his comment; however, for this item the RadSTARD elaboration document states that clearly understanding to whom the study results applied is the most relevant question when assessing the methodology of accuracy in a radiology study. In other words, details about participants with the disease entity versus those without the disease need to be described as this enables future extrapolation of study findings for clinical practice (Black, 1990). Although the word "prevalence" was not used per se, it is implied that if the study population is described this would include a description of the disease prevalence.

RadSTARD item 4: Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard

To which he added, "I think RadSTARD item 19 should be brought into this or directly after this as it makes more sense to include it in the TESTS rather than the RESULTS section."

RadSTARD item 19: The time difference between the index test and reference test and details of any other treatments provided between the two tests should be provided

I concur with this suggestion.

RadSTARD item 10: Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies

The Fellow commented that the description of the index test must include the study parameters (for example sequences used in MR with TE and TR) as well as the model and make of the machine. There are considerable differences in the different generations of CTs as well as differences among different vendors. This would enable a reader to ascertain if this could be applicable in his practice depending on the equipment available.

This item was vetted with the radiological experts when RadSTARD was being developed and it did not meet consensus.

RadSTARD item 14: The training and number of investigators should be described including any extra training for new techniques

For this item the Fellow offered several opinions:
"In case of operator dependant modalities (ultrasound or fluoroscopy) it should be clarified if all the studies were done by one investigator, or if they were done randomly by multiple technicians. This would help analyse source of inter-observer discrepancies in case of operator dependant modalities."

"In case of multiple readers, level of experience of each reader, if the readers were blinded to the reports of each other, and how the differences were handled. Were they compared to reference standard and reported separately or resolved by consensus?"

"Also, if two studies were interpreted, for example CT and MR enterography, it should be clarified if the same observer read both of the studies or there were different observers. If the same observer read them, was there a time interval difference between the first study read and the second study? There is potential of recall bias if the same observer reads both CT and MR for the same patient with a short time between each other even if the studies are anonymized. A 6 week interval is suggested."

These comments were addressed by RadSTARD as per item 14, 15 and 19.

RadSTARD item 14: With respect to interpretation of the index test and reference standard, the training and number of investigators should be described as recommended by STARD and RadSTARD. As reader variability was found within the field of diagnostic imaging, the level of training could impact whether similar results can be found in clinical practice of varying degrees of experience (Bossuyt et al., 2003b, Riegelman, 2013). In addition, the RadSTARD recommends including a description of any extra training for new techniques.

The provision of additional factors added to a checklist for distinct classes of diagnostic imaging tests has been recommended. For example, when assessing the accuracy of medical imaging modalities, certain items on the checklists may differ depending on their technological stage of development when the study was performed. Likewise, clinical context or variability amongst interpreters may be more significant pending the study stage when aspects of technical parameters are being established for new imaging modalities (Gatsonis, 2003).

RadSTARD 15: Whether the interpreters of the index test or reference standard were blinded to the results of either prior study results, or clinical information, should be described.
Essentially, if those interpreting results of the index test were not blinded to the results obtained from the reference standard, the interpretation of those results could be inflated affecting the overall extent of diagnostic accuracy (Bossuyt et al., 2003a). Both STARD (item 11) and the RadSTARD are in agreement with this recommendation.

Determining how patients are selected for evaluation influences how well the results reflect the true accuracy of the test for the study population. Whether accuracy is reported as sensitivity and specificity with a numerical value, or implicitly through diagnostic imaging interpretation, almost all processes of diagnostic evaluation are biased to some degree. To avoid test review bias radiologists should be blinded to other contributory findings, as this will ensure that the diagnosis was made entirely on the study results. Such interpretation will minimize the possibility of overestimating the degree of sensitivity and specificity (Black, 1990).

As an example when a radiologist attempts to differentiate hemangioma of the liver versus metastatic disease via magnetic resonance imaging, certain diagnostic criterion must be met. If prior knowledge of the diagnosis was known, subtle ambiguities such as the contour and shape of the liver, as well as relative signal intensity on magnetic resonance imaging, could impact the radiologists interpretation of the lesion (Black, 1990). In other words, if the interpreter knows the results for the index test and other clinical results when interpreting the reference standard for the study cohort this could inflate the degree of diagnostic accuracy. By thoroughly and accurately reporting diagnostic accuracy studies, the incidence of bias can be minimized, and generalizability of results, assessed (Smidt et al., 2005).

RadSTARD 19: **The time interval that it took to perform the index test and the reference standard is item number 17 of the STARD recommendation tool.**
The reason for including this information is that if there was a time lapse in the patient's condition, it could change and affect the target condition being studied (Bossuyt et al., 2003a). In addition, the STARD tool recommends reporting any treatment or investigations that may have occurred between the performance of the index test and the reference standard or estimates and impact the overall results. It is

recommended that the index test and reference standard be performed within a short interval of time, since the administration of any intervention or treatment between the two tests may affect the ability of the second test to diagnose the disease under study. Therefore, information on the assignment of how patients were recruited with respect to the time interval for receiving or conducting the index test and reference standard should be included when conducting and/or interpreting diagnostic accuracy studies (Riegelman, 2000).

RadSTARD also recommends reporting the time interval between the index test and reference standard and the provision of any additional treatment for the same reasons as recommended by STARD. However, when radiologists are interpreting results of the two tests, reading-order bias can also occur. Suppose a patient undergoes two tests - treatment A and treatment B. If they are read by the same interpreter, the images read last (treatment B) by radiologists will be interpreted more accurately than the images retained from treatment A (Obuchowski, 2003b).

When conducting diagnostic accuracy studies, one typically interprets the results of the index test after the performance of the reference standard where the method of interpretation is done by two separate readers. This type of reading-order bias is mentioned as it can occur in other types of radiology research (Obuchowski, 2003b). As a result, the RadSTARD recommends that the reader be cognizant that this type of bias did not occur.

Assessing the diagnostic accuracy for the index test involves verifying the results from the index test to that of the reference standard for each participant studied. If not all the participants are verified, or some of the participants are verified by additional reference standards, this can result in verification bias. Partial verification bias occurs when not all study participants undergo the reference standard resulting in an over-estimation of diagnostic accuracy (Leeflang et al., 2009).

RadSTARD item 20: Adverse events should be reported for either the index test or reference standard

The Fellow recommended the following points: “Along with adverse effects, data on image quality needs to be provided. It needs to be stated if any of the studies were non-diagnostic or had low image quality and needed to be repeated or call back of the patient. For example, MRI is excellent for diagnosing pancreatitis, but it is impractical as it is plagued by respiratory motion artefacts in at least 30% of sick patients”.

I do not concur with his comment. Adding information on image quality is addressed by item 11 of RadSTARD. The following is an example of the detail that should be provided when reproducing the analysis of functional MRI (fMRI). In an effort to ensure that the reader comprehends the model being used, it is pertinent to describe the approach in detail. fMRI analysis is typically reported using the general linear model (GLM) whereby there are various differences between the models that should be defined. Most of these differences are apparent upon analysis with the software packages. Therefore, rationale such as how the covariance structure was modelled should be thoroughly described for the software parameters utilized. When comparing the results from individual studies, provision of the precise statistical tests utilized to draw inferences should also be included (Poldrack et al., 2008).

“It may also be worth mentioning the agreeable side effects. For example, MR enterography needs administration of intravenous buscopan, which may impair vision for a short while, and the patient may not be able to drive the car back home or would need to bring along a driver. Polyethylene glycol administration may result in acceptable but discomforting diarrhea.”

This was an interesting point raised however; I do not think it would be described in the literature.

RadSTARD item 23: Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability including > 3 observers with variable expertise

The Fellow commented, “I think having >3 observers for each study is too strict and not feasible in every case. It can be more a guideline, which is difficult to follow. I think it needs to be rephrased as “should provide data on inter-observer variability among readers with differing experience levels.” The RadSTARD recommends that whenever an observer reports the results of the test, there is a potential for either inter-observer or intra-observer variability to occur. For example, if two radiologists provided different interpretations for the same digital image it is called inter-observer variation. Whereas if an interpreter reported any significant difference between two different readings throughout the day, this would be referred to as intra-observer variation (Riegelman, 2000). RadSTARD recommends including > 3 observers of variable experience when interpreting the study results.

RadSTARD item 24: Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility

The Fellow commented, “I am not quite clear regarding the difference between 23 and 24. Point 23 was about inter-observer variability in appreciating imaging findings. What is test reproducibility? Do you mean the difference in acquisition of USG images by 2 technologists? If so, how is kappa calculated?”

I believe this was only a point of clarification for the Fellow as item 24 of RadSTARD states both STARD and RadSTARD recommend reporting the estimates of test reproducibility. This is commonly conducted with the coefficient of variation and Kappa statistics (Bossuyt et al., 2003b). Reproducibility of study results may also be described as reliability meaning that the same results would be achieved if the test were repeated again. Although the STARD criterion recommend reporting quantitative assessments for reproducibility it does not require reporting the conduct of the methods utilized for how the estimates of reproducibility were performed (Riegelman, 2000).

This item corresponds to item 24 of the STARD. Reporting estimates of reproducibility is commonly conducted with the coefficient of variation and kappa statistics.

Additional comments for consideration from the same Fellow included:
Cost consideration of the index study – it should be clarified if the index test would significantly affect healthcare costs, and if it would, will the advantages be worth it. For example, MRI is more sensitive than mammography, but it is impractical as a screening tool.

I think this cost-benefit analysis is important, but including it in a checklist of items to report based on the conduct of a diagnostic accuracy study may not be appropriate. Instead, I would suggest conducting a cost-benefit analysis study where the costs and cost-benefit ratios are properly calculated.

Time consideration - it should be clarified, how long the study takes. CT takes a few minutes whereas MR may take longer which may potentially delay management. This would enable clarification if any abbreviated or faster MR sequences with shorter time were used.

This was previously addressed as per item 19 of RadSTARD.

If the information from the index test was used in the decision-making process for the clinical team during the course of the study. If so, in what percentage of the study population was management altered due to the imaging findings in the reference study? For example, CT may detect multiple additional tiny collections in pancreatitis when compared to USG, but management would not have differed.

This item is addressed with item 12 of RadSTARD. RadSTARD recommends reporting any modifications that occurred during the conduct of a radiology diagnostic accuracy study. The STARD tool does not describe this item. As there are reportedly many potential sources of bias in radiologic studies, any modifications during the conduct of a radiology diagnostic accuracy study could bias any aspect of the study. For example, interpreting the result of the index test must be done without knowledge of the results from the reference standard. Similarly, the reference standard results must be interpreted without knowing the results from the index test (Obuchowski, 2003b).

Quote previous similar studies in the literature and how the present study differed from them in regards to the methodology and technique. For example, a study done on 64-detector CT versus 4-slice CT would have differing interpretations. This type of information is typically already provided within the body of the article in either the introduction or discussion section.

Lastly, the Fellow commented, “I need to clarify here that the present form of the RADSTARD that you have most painstakingly created is very good and needs widespread acceptance. My criticism is for the sole purpose of constructive feedback.”

5.19 Findings Conclusion

In summary, the validation phase of the RadSTARD tool between the radiology residents and Fellows was completed via triangulation of data as both quantitative and qualitative analysis was completed. The results found no significant statistical difference between the two groups as per the Mann-Whitney and chi-square analysis. Likewise, both physician groups indicated that they found RadSTARD increased their level of confidence when interpreting the diagnostic article by Henes et al. (2012). Concomitantly, when combined, 96% of the two physician groups indicated they would use the tool again.

Therefore, these results may be interpreted as rendering generalizability in that utilization of the new tool did not result in a central tendency between the two physician groups despite their level of training in radiology. This means that there was no discrepancy or statistical difference found in the results between the radiology residents and Fellows scores despite the difference in their level of training. Both groups found the RadSTARD tool and elaboration document to be of benefit to them when interpreting the literature. RadSTARD is a reliable tool that can be used to validate the results of diagnostic accuracy studies specific to radiology.

The next chapter will address recommendations for use of the RadSTARD tool as well as suggestions on how to increase adherence of its use.

6. Chapter 6 Conclusions

6.1 Reporting Tools and Impact on Practice

When we look at the inherent benefits when using reporting tools for reporting or interpreting the literature such as diagnostic accuracy studies, I think it's not difficult to see why they should be used. After all, reporting tools such as the STARD provide researchers guidance on what to include in their diagnostic accuracy articles that are being sent for publication (Bossuyt et al., 2003b). Perhaps the greatest benefit to reporting tools is that they reduce the incidence of inherent bias that can occur when the literature is poorly reported (Moher et al., 2014c). In addition, adherence to reporting tools also reduces the tendency to spin the results which occurs when authors over-inflate their results or leave out data that should have been included in the final analysis (Ochodo et al., 2013).

When research be it diagnostic accuracy, randomized controlled trials or retrospective reviews are reported accurately, without bias results in a cascade of benefits. The stakeholders that benefit include the researchers, other clinicians and trainees reading their results and finally the impact that such knowledge translation has on patient care. If research is not reported accurately, those interpreting the results have to decide whether or not the results as reported would impact their practice. In other words, based on their interpretation of the literature, would it change how they practice medicine?

With more than 70,000 articles being published each day, clinicians are faced with a plethora of literature to sift through (Moher et al., 2010b). Pending their area of expertise, not all physicians and trainees have the knowledge to understand research methodology and statistical results as they lack training in epidemiology and research methods. Hence, reporting tools are beneficial for researchers and clinicians to use when reporting and interpreting the literature but this is with the caveat that they are adhered to. The main purpose of my doctorate was to study the current STARD tool with radiological experts to determine if a revised tool specific to radiology would benefit the radiology trainees when interpreting the literature. The main difference in the development of this tool as compared to every other reporting tool is that this tool was developed by the stakeholders that would benefit the most.

6.2 Recommendations for the RadSTARD

Given my tenure in research, I am fortunate in that my work environment has rendered me the opportunity to advance my knowledge in research. For example, in my role as a senior research professional within the department of radiology, my experiential research knowledge was significantly augmented with the completion of my MSc in research. At that time, my final thesis focused on the adherence to reporting tools when reporting the results of diagnostic accuracy studies. Overall, I found that reporting tools such as the STARD were slowly being adopted by the various authors and publishers but there was certainly a gap with respect to adherence. More importantly, I found that there were items from the STARD list that were not relevant to radiology diagnostic accuracy studies and that there was a need to study this further. Therefore, the results from my master's thesis lead to my desire to create new knowledge which was the development of a new reporting tool called the RadSTARD which was specific to radiology diagnostic accuracy studies.

Collectively, I collaborated with most radiologists and their trainees within my department on this study. I chose participatory action research to develop new knowledge – a revised tool (RadSTARD) based on the current STARD model. My methodology involved acting, planning and reflection for each action research cycle, which in many respects were synonymous with the cycles of the Delphi technique. Likewise, the validation phase, which involved testing the new tool, was also completed in two separate rounds of cycles that also complemented the methods employed by action research and the critical paradigm. Reflecting on practice and learning was a key element in the results obtained that lead to each successive cycle. For instance, in the development phase of the new tool, the radiological experts were required to reflect on their professional practice and knowledge pertaining to diagnostic accuracy in order to provide their responses on the items they thought should remain or be added to the new tool.
In order to validate the new tool, radiology residents and Fellows also needed to reflect on their prior research experience and radiology knowledge to offer their responses as they rated their level of confidence in interpreting a radiology diagnostic accuracy article with the RadSTARD tool. This led to the final cycle of this action research study, as the radiology residents and Fellows provided their feedback on the usefulness of the RadSTARD and whether they would use the tool again. The response rate from the radiology residents and Fellows for this cycle was again good and overall; collectively they found the RadSTARD increased their confidence in interpreting the literature whereby 96% reported they would use the tool again.

By conducting this study with my department, it has resulted in raising awareness on the importance of reporting guidelines. However, in an effort to keep this momentum going it is critical that dissemination of this new tool is done and the tool is adhered to. As discussed previously, one of the most difficult challenges for those who have developed reporting guidelines such as the STARD is ensuring adherence of the tool when authors are submitting their articles for publication (Wilczynski, 2008, Smidt et al., 2005). This next section will discuss recommendations for dissemination and suggestions to hopefully increase the utilization and adherence rate for the RadSTARD tool.

6.3 Conclusions and Going Forward

Ideally, reporting tools such as the RadSTARD would be beneficial at all stages when conducting diagnostic accuracy research, including the development of the protocol, rating the overall quality of diagnostic accuracy research, and lastly as a tool to interpret the literature. Adherence to such tools is integral to successfully conducting diagnostic accuracy research that is deemed methodologically sound and lacks inherent biases (Moher et al., 2010b). Yet, developing and promoting tools such as the RadSTARD will more than likely be difficult, given the fact that most journal editors do not enforce the use of reporting guidelines (Kunath et al., 2012).

I firmly believe that the research process including properly reporting research outcomes needs to begin in medical school. If medical students were exposed to research methods and epidemiology during their medical school training, the use of reporting tools could also be incorporated into their curriculum. By introducing the use of reporting tools early on in their academic career, I am confident the uphill battle of trying to promote the use of reporting guidelines may be a thing of the past. As discussed in my review of the literature Chapter 2, there lies a distinction between journals that are so called “adopters” of reporting guidelines versus those that are not (Smidt et al., 2006:796).

As per the EQUATOR network, they have identified certain journals as those that recommend authors refer to reporting guidelines. They include BMJ, Plos Medicine, BJOG (British Journal of Obstetrics), and Annals of Emergency Medicine. Moher et al. (2014c) have recommended that once a guideline has been piloted this is an excellent segue to publication of the reporting guideline. As the RadSTARD was validated with the radiology residents and Fellows, this would serve as piloting of the tool. Therefore, the next step would be to submit the RadSTARD tool, elaboration document, and 2-page summary to several relevant journals. Moher et al. (2014c) have suggested that perhaps the best way of achieving readership about a new guideline is to make it available in open access journals. Therefore, the first journal that I would submit the RadSTARD tool would be Plos Medicine as it is an open-access journal that already endorses the use of reporting guidelines. Naturally, I would submit the RadSTARD to the journal, Radiology as well as the tool is specific to the reporting of radiology diagnostic accuracy studies. Concomitantly, I will present the results from my research at academic rounds within our department, plus I will submit an Abstract to RSNA in spring 2016. RSNA is the largest radiological conference in the world. Providing I am accepted, this will result in dissemination of the RadSTARD on a global platform.
In disseminating the results from my research, I will be recommending that the RadSTARD be used when conducting, reporting and interpreting diagnostic accuracy research specific to radiology. However, in order to invoke a change in practice, exposure to reporting tools should be done in medical school. As

the old adage goes, “publish or perish.” Physicians in academic teaching centres are fully aware of their requirements to publish if they hope to attain academic positions at the university level.

Publishing is not an easy task, overall. In addition, poor reporting practices can seriously compromise the usefulness and reliability of study findings (Moher et al., 2014a). Reporting guidelines can thus help authors improve the quality of their reporting, and their use may even enhance the rate that articles are accepted for publication. When research is reported accurately and reporting guidelines are followed, many benefits will ensue. First and foremost, physicians benefit, as the information from the published study allows them to assess the methodology and overall generalizability of study findings (Moher et al., 2014a). Naturally, such knowledge is then conveyed to the patient, and this is where the practice turn is truly engaged – through one properly reported article at a time (Schatzki et al., 2000).

7. Chapter 7

7.1 Personal Reflection – Introduction

When I reflect back on my learning throughout my professional doctoral study, one of the most important revelations for me was finding a way to use action research in a scientific environment. As stated by McNiff (1995), action research methodology allows researchers to be in control of their research and its context. The underlying theme within healthcare research is that action research is a vehicle that affords collaboration, review, discussion and critical reflection on each aspect of the study (Koshy et al., 2011). Within the radiology department, we routinely conduct research whereby the methodology is unique to each proposal. However, the research methodologies utilized are commonly just a description of how the study will be conducted, such that research paradigms and one's positionality are not routinely described.

Therefore, critical reflection was required to determine which paradigm I believed would provide the foundation for my action research study. Action research prefers "a dialectical view of rationality," while rejecting the "positivism notions of rationality, objectivity and truth" (Koshy et al., 2011:36). For those involved in this action research project, with each iterative cycle they were continually reflecting and evaluating based on their own practice and research theories. These fundamental theories are akin to what Schön defines as the "reflective practice," meaning that practitioners are in control of the knowledge as they are engaged in the process. Thus, the role of "critical reflection" is an essential feature of action research (Koshy et al., 2011:37).

Due to the openness of action research, it is a methodology that stands in stark contrast to our typical hypothesis-driven research (McNiff, 1995). Burns (2005) identifies how novice action researchers may initially struggle with the "inherent tension in the terms, action and research" – as the two words conjoined "lie as uneasy bedfellows" (Burns, 2005:59-60). Having worked in research for more than two decades, I concur with this sentiment. For example, I remember the looks I received from some of the radiology residents and Fellows when I introduced my study to them, explaining that it was an action research study. When the terms "needs assessment" and "Delphi technique" were expressed, however, this somewhat alleviated their concerns. I knew they were interested in the whole process of developing a new tool and then validating was at times quite messy. As Cook (2009) describes, this is one of the main purposes of action research. Perhaps the best way of describing my professional learning and reflection on this methodology is the following quote: "My experience of action research is that is difficult to grasp or explain the concept until one is in the process of doing it" (Burns, 2005:60).

I was grateful for the flexibility that the action research cycles allow, as it allowed me the time to reflect, act and plan subsequent cycles (Koshy et al., 2011). In order to conduct my research within a scientific arena, it required a great deal of methodological planning in that I needed to develop methods that would foster maximum engagement with those who collaborated with me. This served both to initially help me develop the tool and also connect with those who agreed to participate during the validation phase of the tool. In my attempt to link action research to the creation of new knowledge, the participatory action research (PAR) methodology worked well, as it allowed me to link participants' existing knowledge to their current practice with the aim of creating a meaningful change (Abraham and Purkayastha, 2012, Brydon-Miller and Maguire, 2009).

7.2 Transdisciplinarity and Knowledge Creation

Based on my tacit and experiential knowledge in radiology research, and my review of the literature, I knew that reporting tools such as the STARD were seldom used or not used to their full potential for radiology diagnostic accuracy studies. This "knowing-in-action" was beneficial, as it provided the framework in which to tackle the problem (Schön, 1987b). In my efforts to address this knowledge fragmentation, I relied on a transdisciplinary approach (Lawrence and Després, 2004). In the "new production of knowledge," published by Gibbons et al. (1994), "Mode 1" refers to knowledge formulated by "traditional disciplinary production" (Balsiger, 2004:407). Concomitantly, "Mode 2" is another type of knowledge that is defined by a transdisciplinary approach. This type of new knowledge creation was denoted as the scientific way of creating new knowledge whereby the transdisciplinarity approach became the way of addressing future research initiatives that focused on addressing issues "that are not circumscribed in any existing disciplinary field" (Balsiger, 2004:407).

As previously described by Moher et al. (2014b), most reporting guidelines are generally not tested or validated, as they are often deemed too complicated. Concomitantly, of the 250 reporting tools that have been previously developed, most were created with senior scientific methodologists and statisticians that met together in collaboration to develop new reporting tools to address poor reporting of the literature. Therefore, developing the methodology to approach the issue of poorly reported radiological diagnostic accuracy studies by creating a revised tool based on the current STARD was challenging and rewarding. This next section will summarize the reflective cycles conducted throughout the development of the new tool and how "Mode 2" knowledge was created via PAR, which required a transdisciplinary approach (Balsiger, 2004:407).

7.3 Transdisciplinary Action Research

Transdisciplinarity is a research standard that was originally developed to deal with research issues pertaining to societal needs. It was particularly beneficial for research problems that were not assigned to any particular disciplinary domain or framework (Balsiger, 2004). Transdisciplinary action research provides the framework required to solve problems based on a concept of action research first developed by Lewin in the 1940s (Stokols, 2006). Action research then became of particular interest to psychologists in the late 1960s and 1970s as psychologists began focusing on developing methods to address societal concerns. Although action research was an accepted methodology, collaboration with others was more difficult than expected. As one particular collaborator was quoted as saying: "A psychologist cannot simply walk in off the street, tell other people what they are doing wrong, walk away, and expect them to change their behavior". Rather, one must work with people…to facilitate the change process" (Stokols, 2006:64).

Action research requires taking action to improve one's practice. It is a self-reflective practice as it involves carefully creating the methods required to conduct research with others (McNiff, 2013). Action research is concerned with collaborative scientific inquiry whereby participants have a vested interest in the project as it is deemed applicable to practice and is context-specific. Action research is also empowering, as participants actively engage in the research process throughout various activities. Participants need to be willing to participate in action research with the caveat that they determine the research to be worthwhile, thereby agreeing to be part of the change process. However, agreeing to participate in action research is often a huge undertaking in that the participants are required to do more than simply answer a few questionnaires, as the research design may be complex whereby subsequent feedback from the participants necessitates changes to the research design (Meyer, 2000).

Whereas transdisciplinary collaborations "force participants out of their disciplinary comfort zones and require their unwavering commitment to sustained and mutually respected communication" (Stokols, 2006:68). Reflecting back on my study, and in particular on the validation phase, I was urging the radiology residents and Fellows to move out of their comfort zones when they rated their level of confidence after reading a radiology diagnostic accuracy article in conjunction with the RadSTARD tool. By subscribing to a transdisciplinary approach via participatory action research, knowledge was produced within the context of our departmental workplace. This type of knowledge is not as constrained as in traditional disciplinary methods (Boud and Tennant, 2006).

7.4 Reflective Practice and Participatory Action Research

As an insider within the department, I was aware of the demands of the radiologists with respect to their professional responsibilities to the hospital and academic duties at the University of Ottawa. Indeed, they are required to demonstrate excellent time management skills to navigate through their plethora of professional obligations. For example, in addition to interpreting diagnostic images within their area of sub-specialty, they are also required to teach residents and Fellows, and provide medical lectures at various times throughout the day. Hence, between all of these obligations (not to mention their own personal lives), many of the radiologists are also involved in research initiatives and publishing their findings. Not all of them are, however, and hence as an insider researcher I knew who would be most likely be willing to participate in my study.

In some respects this may be construed as biased, as I cherry-picked my collaborators; however, choosing a radiological expert who did not conduct research or publish seemed to be counter-productive. As the study advanced and upon deeper reflection, I was trying to strike a fine balance as an insider collaborating with those who perceived me as an outsider (Schön, 1987b, Costley et al., 2010). Although I was fully aware of the time commitments and obligations the radiological experts had to their professional practice given the time it took for them to complete the needs assessment and Delphi rounds, I often felt like an outsider (Smith et al., 2010).

I allowed myself time to process this feeling of angst as I realized I was learning. I researched ways to handle this challenge as others have faced similar challenges with PAR (Kindon and Elwood, 2009). I have learned thus far that the true benefit of action research is the overall flexibility that this methodology affords the researcher. I also needed to be flexible and develop “soft skills” to navigate the challenges of the PAR stages while we developed the tool. Examples of soft skills are facilitating and managing group dynamics, and having an awareness of one’s “emotional intelligence,” which encompasses critical reflection, social skills and empathy (Kindon and Elwood, 2009:25). Other skills include learning how to negotiate and manage conflicting situations. Upon critical reflection in action, I used many of these soft skills in my previous career endeavors; however, I became much more aware of these actions as I conducted my doctoral study. Consequently, I realized that my reflections on these actions from previous experience remained somewhat dormant until I consciously attended to them. Indeed, reflecting on my research study closed the gap between me, the researcher, and those who collaborated with me. By removing this gap, my relationship with my collaborators resulted in “a sense of power, involvement and agency” (Etherington, 2004:32).

7.5 Action Research and Rigor

Within my dissertation, I have discussed why this study was needed, and what the overall purpose of the study was. As described by Wardlaw et al. (2012), diagnostic imaging should be evaluated as rigorously as pharmacological interventions, as this will improve the quality of healthcare provided to patients. Ochodo et al. (2013) stressed the significance of transparency required when reporting evaluations of studies that assess diagnostic tests or medical tests as they can be fraught with inherent biases. Essentially, my objective was to develop a high-quality systematic tool that could be validated within our department through participatory action research, a methodology that encourages group participation and community-based scientific enquiry.

Perhaps one of the most notable concerns of action research raised by others is its lack of rigor and validity. Action research is criticized for not allowing objectivity into one's own practice when conducting the research project. However, this concern is managed by acknowledging my epistemological stance from the onset. Arguably, this could be a concern within a positivistic environment; however, our group was accustomed to utilizing questionnaires to conduct research where input from all is required to obtain the data required for analysis. Indeed, Koshy et al. (2011:33) advises that a "validation group” be established to share the data with before giving it to the working group.

During each phase of my study, I presented the research findings and discussed each stage of the process with my colleagues and mentor before providing feedback to the working group. The methods chosen for this action research study of gathering, reflecting and analyzing data within our department were rigorous. In addition, validation of the newly developed RadSTARD tool with the radiology residents and Fellows rendered triangulation as mixed methods were utilized. This was essential for ensuring the overall quality of the research methods and data collection, enabling it to be considered robust, transparent and without bias.

Equally concerning as the establishment of rigor in action research is the outcome of knowledge. This indeed is an interesting question. Is action research conducted in a cyclical fashion with the sole purpose of validating mixed methods? And who benefits from this research and knowledge? It is equally important that the knowledge gained from action research be placed into existing frameworks of epistemology, and one way of ensuring this is through the dissemination of the findings (Gray, 2013).

7.6 Closing Notes and Future Directions

The results of this participatory action research led to the development of a new reporting tool for radiologists and their trainees to use when interpreting radiology diagnostic accuracy studies. The original STARD tool was recommended for use only when reporting on the results of diagnostic accuracy, as Bossuyt (2009a) did not prescribe its use in any other context. I disagree with this recommendation. Rather, I am recommending that the RadSTARD be used by radiologists and their trainees at all stages of the research process, including protocol development.

However, before this can be achieved I am cognizant of the concern over adherence to the RadSTARD, given the lackluster adherence to other similar reporting tools. Certainly, publishing is necessary but this is only one component of promoting literature that is deemed methodologically sound and added to the repository of evidence-based medicine.

I do believe that creating the RadSTARD within our radiology department by those who will benefit the most by its use has certainly captured the interest of those who agreed to participate in the research. Concomitantly, given that this new tool was created by radiologists for radiologists is also quite novel. Through dissemination via publication, and presenting at local and national academic conferences, I am optimistic that this new knowledge will aid radiologists in reporting and interpreting radiology diagnostic accuracy studies that will impact their practice for generations to come.

8. References:

ABRAHAM, M. & PURKAYASTHA, B. 2012. Making a difference: Linking research and action in practice, pedagogy, and policy for social justice: Introduction. *Current Sociology,* 60, 123-141.

AFIFY, M. F. 2008. Action research: Solving real-world problems. *Tourism and Hospitality Research,* 8, 153-159.

AL-SULTTAN, F. M., FRAGKOS, K. C., BOGDANOS, D. P. & FORBES, A. 2012. Tu1224 A Systematic Review and Meta-Analysis of Pancreatic Autoantibody's (Pab) Diagnostic Accuracy vs Standard Diagnosis in Patients With Inflammatory Bowel Disease. *Gastroenterology,* 142, S-778-S-779.

ALSHAMARI, M., NORRMAN, E., GEIJER, M., JANSSON, K. & GEIJER, H. 2015. Diagnostic accuracy of low-dose CT compared with abdominal radiography in non-traumatic acute abdominal pain: prospective study and systematic review. *European Radiology*, 1-9.

ALTMAN, D. G. & MOHER, D. 2014. Importance of Transparent Reporting of Health Research. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

APPLEGATE, K. E. & CREWSON, P. E. 2002. An introduction to biostatistics. *Radiology,* 225, 318-322.

ASTIN, M. P., BRAZZELLI, M. G., FRASER, C. M., COUNSELL, C. E., NEEDHAM, G. & GRIMSHAW, J. M. 2008. Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the diagnostic performance of imaging techniques. *Radiology,* 247, 365-373.

BACHMANN, L. M., TER RIET, G., WEBER, W. E. & KESSELS, A. G. 2009. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. *Journal of clinical epidemiology,* 62, 357-361. e2.

BAILEY, B. & AMRE, D. K. 2005. A toxicologist's guide to studying diagnostic tests. *Clinical Toxicology,* 43, 171-179.

BALSIGER, P. W. 2004. Supradisciplinary research practices: history, objectives and rationale. *Futures,* 36, 407-421.

BANSAL, G. J. & YOUNG, P. 2015. Digital breast tomosynthesis within a symptomatic “one-stop breast clinic” for characterization of subtle findings. *The British Journal of Radiology,* 88, 20140855.

BARBOUR, R. S. 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ,* 322, 1115-1117.

BARDOU, D., BOURSIER, J., CARTIER, V., LEBIGOT, J., MICHALAK, S., OBERTI, F., FOUCHARD-HUBERT, I., ROUSSELET, M. C., AUBE, C. & CALÈS, P. 2013. FIRST INTENTION-TO-DIAGNOSE COMPARISON OF ARFI AND FIBROSCAN IN CHRONIC LIVER DISEASES. *Journal of Hepatology,* 58, S7.

BERTENS, L. C. M., BROEKHUIZEN, B. D. L., NAAKTGEBOREN, C. A., RUTTEN, F. H., HOES, A. W., VAN MOURIK, Y., MOONS, K. G. M. & REITSMA, J. B. 2013. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Medicine,* 10, e1001531.

BLACK, W. C. 1990. How to evaluate the radiology literature. *American Journal of Roentgenology,* 154, 17-22.

BOHTE, A. E., DE NIET, A., JANSEN, L., BIPAT, S., NEDERVEEN, A. J., VERHEIJ, J., TERPSTRA, V., SINKUS, R., VAN NIEUWKERK, K. M. & DE KNEGT, R. J. 2014. Non-invasive evaluation of liver fibrosis: a comparison of ultrasound-based transient elastography and MR elastography in patients with viral hepatitis B and C. *European Radiology,* 24, 638-648.

BOSSUYT, P. M. 2008a. STARD Statement: Still Room for Improvement in the Reporting of Diagnostic Accuracy Studies 1. *Radiology,* 248, 713-714.

BOSSUYT, P. M. 2009a. Diagnostic accuracy reporting guidelines should prescribe reporting, not modeling. *Journal of Clinical Epidemiology,* 62, 355-356.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L., LIJMER, J. G., MOHER, D., RENNIE, D. & DE VET, H. C. 2015a. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology*, 151516.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L., LIJMER, J. G., MOHER, D., RENNIE, D. & DE VET, H. C. 2015b. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology,* 277, 826-832.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., LIJMER, J. G., MOHER, D., RENNIE, D., DE VET, H. C. & STANDARDS FOR REPORTING OF DIAGNOSTIC, A. 2003a. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Radiology,* 226, 24-8.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., MOHER, D., RENNIE, D., DE VET, H. C., LIJMER, J. G. & STANDARDS FOR REPORTING OF DIAGNOSTIC, A. 2003b. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem,* 49, 7-18.

BOSSUYT, P. M. M. 2008b. Interpreting Diagnostic Test Accuracy Studies. *Seminars in Hematology,* 45, 189-195.

BOSSUYT, P. M. M. 2009b. Diagnostic accuracy reporting guidelines should prescribe reporting, not modeling. *Journal of clinical epidemiology,* 62, 355-356.

BOSSUYT, P. M. M. 2014. STARD (STAndards for Reporting of Diagnostic Accuracy Studies). *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

BOUD, D. & TENNANT, M. 2006. Putting doctoral education to work: challenges to academic practice. *Higher Education Research & Development,* 25, 293-306.

BOURSIER, J., DE LEDINGHEN, V., POYNARD, T., GUÉCHOT, J., CARRAT, F., LEROY, V., WONG, G. L.-H., FRIEDRICH-RUST, M., FRAQUELLI, M. & PLEBANI, M. 2015. An extension of STARD statements for reporting diagnostic accuracy studies on liver fibrosis tests: the Liver-FibroSTARD standards. *Journal of Hepatology,* 62, 807-815.

BRADBURY, H. & REASON, P. 2006. *Handbook of Action Research*, Sage.

BRAT, R., YOUSEF, N., KLIFA, R., REYNAUD, S., AGUILERA, S. S. & DE LUCA, D. 2015. Lung Ultrasonography Score to Evaluate Oxygenation and Surfactant Need in Neonates Treated With Continuous Positive Airway Pressure. *JAMA Pediatrics,* 169, e151797-e151797.

BREALEY, S. & SCALLY, A. J. 2008. Methodological approaches to evaluating the practice of radiographers' interpretation of images: A review. *Radiography,* 14, e46-e54.

BRELL, M., IBÁÑEZ, J. & TORTOSA, A. 2011. O6-Methylguanine-DNA methyltransferase protein expression by immunohistochemistry in brain and non-brain systemic tumours: systematic review and meta-analysis of correlation with methylation-specific polymerase chain reaction. *BMC Cancer,* 11, 35.

BRODIE, P. & IRVING, K. 2007. Assessment in work-based learning: investigating a pedagogical approach to enhance student learning. *Assessment & Evaluation in Higher Education,* 32, 11-19.

BRUCHER, N., VIAL, J., BAUNIN, C., LABARRE, D., MEYRIGNAC, O., JURICIC, M., BOUALI, O., ABBO, O., GALINIER, P. & SANS, N. 2015. Non-contrast-enhanced MR angiography using time-spin labelling inversion pulse technique for detecting crossing renal vessels in children with symptomatic ureteropelvic junction obstruction: comparison with surgical findings. *European Radiology*, 1-8.

BRYDON-MILLER, M. & MAGUIRE, P. 2009. Participatory action research: Contributions to the development of practitioner inquiry in education. *Educational Action Research,* 17, 79-93.

BUDOVEC, J. J. & KAHN, C. E. 2010. Evidence-Based Radiology: A Primer in Reading Scientific Articles. *American Journal of Roentgenology,* 195, W1-W4.

BUDOVEC, J. J. & KAHN JR, C. E. 2010. Evidence-based radiology: a primer in reading scientific articles. *American Journal of Roentgenology,* 195, W1-W4.

BUNNISS, S. & KELLY, D. R. 2010. Research paradigms in medical education research. *Medical Education,* 44, 358-366.

BURCH, J., SOARES-WEISER, K., ST JOHN, D., DUFFY, S., SMITH, S., KLEIJNEN, J. & WESTWOOD, M. 2007. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. *Journal of Medical Screening,* 14, 132-137.

BURNS, A. 2005. Action research: An evolving paradigm? *Language Teaching,* 38, 57-74.

CANTISANI, V., MACERONI, P., D'ANDREA, V., PATRIZI, G., DI SEGNI, M., DE VITO, C., GRAZHDANI, H., ISIDORI, A. M., GIANNETTA, E. & REDLER, A. 2015. Strain ratio ultrasound elastography increases the accuracy of colour-Doppler ultrasound in the evaluation of Thy-3 nodules. A bi-centre university experience. *European Radiology*, 1-9.

CARP, J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage,* 63, 289.

CASSINOTTO, C., LAPUYADE, B., AÏT-ALI, A., VERGNIOL, J., GAYE, D., FOUCHER, J., BAILACQ-AUDER, C., CHERMAK, F., LE BAIL, B. & DE LÉDINGHEN, V. 2013. Liver fibrosis: noninvasive assessment with acoustic radiation force impulse elastography—comparison with FibroScan M and XL probes and FibroTest in patients with chronic liver disease. *Radiology,* 269, 283-292.

CENGIZ, M., SENTÜRK, S., CETIN, B., BAYRAK, A. H. & BILEK, S. U. 2014. Sonographic assessment of fatty liver: intraobserver and interobserver variability. *International Journal of Clinical and Experimental Medicine,* 7, 5453-5460.

CHANTEAU, S., DARTEVELLE, S., MAHAMANE, A. E., DJIBO, S., BOISIER, P. & NATO, F. 2006. New rapid diagnostic tests for Neisseria meningitidis serogroups A, W135, C, and Y. *PLoS Med,* 3, e337.

CHAVEZ, C. A., SKI, C. F. & THOMPSON, D. R. 2014. Psychometric properties of the Cardiac Depression Scale: A systematic review. *Heart, Lung and Circulation,* 23, 610-618.

CHEN, W., MO, J.-J., LIN, L., LI, C.-Q. & ZHANG, J.-F. 2015a. Diagnostic value of magnetic resonance cholangiopancreatography in choledocholithiasis. *World Journal of Gastroenterology: WJG,* 21, 3351.

CHEN, W., XING, W., PENG, Y., HE, Z., WANG, C. & WANG, Q. 2013. Cerebral Aneurysms: Accuracy of 320–Detector Row Nonsubtracted and Subtracted Volumetric CT Angiography for Diagnosis. *Radiology,* 269, 841-849.

CHEN, X., YANG, Y., GAN, W., XU, L., YE, Q. & GUO, H. 2015b. Newly Designed Break-Apart and ASPL-TFE3 Dual-Fusion FISH Assay Are Useful in Diagnosing Xp11. 2 Translocation Renal Cell Carcinoma and ASPL-TFE3 Renal Cell Carcinoma: A STARD-Compliant Article. *Medicine,* 94, 1-8.

CHIESA, C., PACIFICO, L., NATALE, F., HOFER, N., OSBORN, J. F. & RESCH, B. 2015a. Fetal and early neonatal interleukin-6 response. *Cytokine*, 76 (1), 1-12.

CHIESA, C., PACIFICO, L., OSBORN, J. F., BONCI, E., HOFER, N. & RESCH, B. 2015b. Early-Onset Neonatal Sepsis: Still Room for Improvement in Procalcitonin Diagnostic Accuracy Studies. *Medicine,* 94.

CHO, N., IM, S.-A., KANG, K. W., PARK, I.-A., SONG, I. C., LEE, K.-H., KIM, T.-Y., LEE, H., CHUN, I. K. & YOON, H.-J. 2015. Early prediction of response to neoadjuvant chemotherapy in breast cancer patients: comparison of single-voxel 1H-magnetic resonance spectroscopy and 18F-fluorodeoxyglucose positron emission tomography. *European Radiology*, 1-12.

CID, J., AGUINACO, R., SÁNCHEZ, R., GARCÍA-PARDO, G. & LLORENTE, A. 2010. Neutrophil CD64 expression as marker of bacterial infection: a systematic review and meta-analysis. *Journal of Infection,* 60, 313-319.

CLIBBENS, N., WALTERS, S. & BAIRD, W. 2012. Delphi research: issues raised by a pilot study. *Nurse Researcher,* 19, 37.

CONNELLY, L. M. 2008. Bland-Altman plots. *Medsurg Nursing : Official Journal of the Academy of Medical-Surgical Nurses,* 17, 175.

COOK, T. 2009. The purpose of mess in action research: building rigour though a messy turn. *Educational Action Research,* 17, 277-291.

COPPUS, S. F., VAN DER VEEN, F., BOSSUYT, P. M. & MOL, B. W. 2006. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. *Fertility and Sterility,* 86, 1321-1329.

CORDONNIER, C., SALMAN, R. A.-S. & WARDLAW, J. 2007. Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting. *Brain,* 130, 1988-2003.

COSSE, C., SABBAGH, C., KAMEL, S., GALMICHE, A. & REGIMBEAU, J.-M. 2014. Procalcitonin and intestinal ischemia: A review of the literature. *World Journal of Gastroenterology: WJG,* 20, 17773.

COSTA, E. A., CUNHA, G. M., SMORODINSKY, E., CRUITE, I., TANG, A., MARKS, R. M., CLARK, L., WOLFSON, T., GAMST, A. & SICKLICK, J. K. 2015. Diagnostic Accuracy of Preoperative Gadoxetic Acid–enhanced 3-T MR Imaging for Malignant Liver Lesions by Using Ex Vivo MR Imaging–matched Pathologic Findings as the Reference Standard. *Radiology*, 142069.

COSTLEY, C., ELLIOTT, G. C. & GIBBS, P. 2010. *Doing work based research: Approaches to enquiry for insider-researchers*, Sage.

COSTLEY, C. G., P. 2006. Researching others: care as an ethic for practitioner researchers. *Studies in Higher Education,* 31, 89-98.

CRIM, J. R., LAYFIELD, L. J., SCHMIDT, R., HANRAHAN, C., LIU, T. & MANCASTER, B. J. 2013. MRI leads to increased false-positive diagnosis of chondrosarcoma. *Skeletal Radiology,* 42, 1044-1045.

DAWSON, G. F. 2012. *Easy Interpretation of Biostatistics: The Vital Link to Applying Evidence in Medical Decisions*, Elsevier Health Sciences.

DUNT, D. & MCKENZIE, R. 2012. Improving the quality of qualitative studies: do reporting guidelines have a place? *Family Practice,* 29, 367-369.

ELIKASHVILI, I., TAY, E. T. & TSUNG, J. W. 2014. The Effect of Point-of-care Ultrasonography on Emergency Department Length of Stay and Computed Tomography Utilization in Children With Suspected Appendicitis. *Academic Emergency Medicine,* 21, 163-170.

ENG, J. 2003. Sample Size Estimation: How Many Individuals Should Be Studied?1. *Radiology,* 227, 309-313.

ENG, J. 2004. Sample Size Estimation: A Glimpse beyond Simple Formulas 1. *Radiology,* 230, 606-612.

ERAUT, M. 2000. Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology,* 70, 113-136.

ERRICO, G., GIORDANO, A. & PALTRINIERI, S. 2012. Diagnostic accuracy of electrophoretic analysis of native or defribrinated plasma using serum as a reference sample. *Veterinary Clinical Pathology,* 41, 529-540.

ETHERINGTON, K. 2004. *Becoming a reflexive researcher: Using our selves in research*, Jessica Kingsley Publishers.

FALLENBERG, E., DROMAIN, C., DIEKMANN, F., ENGELKEN, F., KROHN, M., SINGH, J., INGOLD-HEPPNER, B., WINZER, K., BICK, U. & RENZ, D. 2014. Contrast-enhanced spectral mammography versus MRI: initial results in the detection of breast cancer and assessment of tumour size. *European Radiology,* 24, 256-264.

FERREIRA, A. & PACHECO, A. 2015. SimTCM: A human patient simulator with application to diagnostic accuracy studies of Chinese medicine. J Integr Med. 2015; 13 (1): 9–19.

FIDALGO, B. M., CRABB, D. P. & LAWRENSON, J. G. 2015. Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach. *Ophthalmic and Physiological Optics,* 35, 315-323.

FLANIGAN, T. S., MCFARLANE, E. & COOK, S. Conducting survey research among physicians and other medical professionals: A review of current literature. Proceedings of the Survey Research Methods Section, American Statistical Association, 2008. 4136-47.

FLETCHER, A. J. & MARCHILDON, G. P. 2014. Using the delphi method for qualitative, participatory action research in health leadership. *International Journal of Qualitative Methods,* 13, 1-18.

FLICKER, L., RITCHIE, C., NOEL-STORR, A. & MCSHANE, R. 2012. Harmonization of reporting standards for studies of diagnostic test accuracy in dementia and related conditions: the STARDdem (STAndards for the Reporting of Diagnostic accuracy studies-Dementia) criteria. *Alzheimer's & Dementia,* 8, P106.

FLORKOWSKI, C. M. 2008. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews,* 29, S83.

FONTELA, P. S., PANT PAI, N., SCHILLER, I., DENDUKURI, N., RAMSAY, A. & PAI, M. 2009. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PLoS One,* 4, e7753.

FOX, M., GREEN, G. & MARTIN, P. 2007. *Doing Practitioner Research*, Sage.

FRATZ, S., SCHUHBAECK, A., BUCHNER, C., BUSCH, R., MEIERHOFER, C., MARTINOFF, S., HESS, J. & STERN, H. 2009. Comparison of accuracy of axial slices versus short-axis slices for measuring ventricular volumes by cardiac magnetic resonance in patients with corrected tetralogy of fallot. *The American Journal of Cardiology,* 103, 1764-1769.

FREEMAN, K., SZCZEPURA, A. & OSIPENKO, L. 2009. Non-invasive fetal RHD genotyping tests: a systematic review of the quality of reporting of diagnostic accuracy in published studies. *European Journal of Obstetrics & Gynecology and Reproductive Biology,* 142, 91-98.

GARDNER, I. A., BURNLEY, T. & CARAGUEL, C. 2014. Improvements are Needed in Reporting of Accuracy Studies for Diagnostic Tests Used for Detection of Finfish Pathogens. *Journal of Aquatic Animal Health,* 26, 203-209.

GATSONIS, C. 2003. Do we need a checklist for reporting the results of diagnostic test evaluations? The STARD proposal. *Acad Radiol,* 10, 599-600.

GEEVASINGA, N., MENON, P., YIANNIKAS, C., HOWELLS, J., KIERNAN, M. & VUCIC, S. 2015. Threshold tracking TMS: A novel diagnostic technique for Amyotrophic Lateral Sclerosis (S24. 005). *Neurology,* 84, S24. 005.

GEFFROY, Y., BOULAY-COLETTA, I., JULLÈS, M.-C., NAKACHE, S., TAOUREL, P. & ZINS, M. 2014. Increased unenhanced bowel-wall attenuation at multidetector CT is highly specific of ischemia complicating small-bowel obstruction. *Radiology,* 270, 159-167.

GEORGANTOPOULOU, C., SIMM, A. & ROBERTS, M. 2008. Transvaginal saline hysterosonography: a comparison with local anaesthetic hysteroscopy for the diagnosis of benign lesions associated with menorrhagia. *Gynecological Surgery,* 5, 27-34.

GLASZIOU, P., ALTMAN, D. G., BOSSUYT, P., BOUTRON, I., CLARKE, M., JULIOUS, S., MICHIE, S., MOHER, D. & WAGER, E. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet,* 383, 267-276.

GODLEE, F. 1994. The Cochrane collaboration. *BMJ : British Medical Journal,* 309, 969-970.

GOEBELL, P. J., KAMAT, A. M., SYLVESTER, R. J., BLACK, P., DROLLER, M., GODOY, G., M'LISS, A. H., JUNKER, K., KASSOUF, W. & KNOWLES, M. A. Assessing the quality of studies on the diagnostic accuracy of tumor markers. Urologic Oncology: Seminars and Original Investigations, 2014. Elsevier, 1051-1060.

GRANT, J. 2002. Learning needs assessment: assessing the need. *Bmj,* 324, 156-159.

GRAY, D. E. 2013. *Doing research in the real world*, Sage.

GRIFFITHS, P., JARVIS, D., MCQUILLAN, H., WILLIAMS, F., PALEY, M. & ARMITAGE, P. 2013. MRI of the foetal brain using a rapid 3D steady-state sequence. *The British journal of radiology,* 86, 20130168.

GRIX, J. 2004. The foundations of research. Basingstoke; New York: Palgrave Macmillan.

GRIX, J. 2010. *The foundations of research*, Palgrave Macmillan.

GUBA, E. G. & LINCOLN, Y. S. 1994. Competing paradigms in qualitative research. *Handbook of Qualitative Research,* 2.

GUSTAVSEN, B. 2001. Theory and practice: The mediating discourse. *Handbook of Action Research: The concise paperback edition*, 17-26.

GUYATT, G., CAIRNS, J., CHURCHILL, D., COOK, D., HAYNES, B., HIRSH, J., IRVINE, J., LEVINE, M., LEVINE, M. & NISHIKAWA, J. 1992. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA,* 268, 2420-2425.

H-ICI, D. O., RIDGWAY, J. P., KUEHNE, T., BERGER, F., PLEIN, S., SIVANANTHAN, M. & MESSROGHLI, D. R. 2012. Cardiovascular magnetic resonance of myocardial edema using a short inversion time inversion recovery (STIR) black-blood technique: Diagnostic accuracy of visual and semi-quantitative assessment. *Journal of Cardiovascular Magnetic Resonance,* 14, 22-22.

HADDOW, L. J., FLOYD, S., COPAS, A. & GILSON, R. 2013. A systematic review of the screening accuracy of the HIV Dementia Scale and International HIV Dementia Scale. *PloS one,* 8, e61826.

HÅKONSEN, S. J., PEDERSEN, P. U., BATH-HEXTALL, F. & KIRKPATRICK, P. 2015. Diagnostic test accuracy of nutritional tools used to identify undernutrition in patients with colorectal cancer: a systematic review. *The JBI Database of Systematic Reviews and Implementation Reports,* 13, 141-187.

HALL, S., LEWITH, G., BRIEN, S. & LITTLE, P. 2008. A review of the literature in applied and specialised kinesiology. *Forschende Komplementärmedizin/Research in Complementary Medicine,* 15, 40-46.

HARRISON, J. K., FEARON, P., NOEL-STORR, A. H., MCSHANE, R., STOTT, D. J. & QUINN, T. J. 2014. Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a general practice (primary care) setting. *Cochrane Database of Systematic Reviews,* 7.

HAUTH, E., HOHMUTH, H., COZUB-POETICA, C., BERNAND, S., BEER, M. & JAEGER, H. 2015. Multiparametric MRI of the prostate with three functional techniques in patients with PSA elevation before initial TRUS-guided biopsy. *The British Journal of Radiology,* 88, 20150422.

HAYNES, R. B. & WILCZYNSKI, N. L. 2004. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *BMJ,* 328, 1040.

HELLEMONS, M. E., KERSCHBAUM, J., BAKKER, S. J., NEUWIRT, H., MAYER, B., MAYER, G., DE ZEEUW, D., LAMBERS HEERSPINK, H. & RUDNICKI, M. 2012. Validity of biomarkers predicting onset or progression of nephropathy in patients with Type 2 diabetes: a systematic review. *Diabetic Medicine,* 29, 567-577.

HENES, F., NÜCHTERN, J., GROTH, M., HABERMANN, C., REGIER, M., RUEGER, J., ADAM, G. & GROSSTERLINDEN, L. 2012. Comparison of diagnostic accuracy of magnetic resonance imaging and multidetector computed tomography in the detection of pelvic fractures. *European Journal of Radiology,* 81, 2337-2342.

HERON, J. & REASON, P. 1997. A participatory inquiry paradigm. *Qualitative inquiry,* 3, 274-294.

HERR, K. & ANDERSON, G. L. 2005. *The action research dissertation: A guide for students and faculty*, Sage.

HEWITT, M., MCPHAIL, M., POSSAMAI, L., VLAVIANOS, P., DHAR, A. & MONAHAN, K. 2011. A meta-analysis of endoscopic ultrasound with fine needle aspiration (EUS-FNA) for diagnosis of solid pancreatic neoplasms. *Gut,* 60, A190-A191.

HING, W., WHITE, S., REID, D. & MARSHALL, R. 2009. Validity of the McMurray's test and modified versions of the test: a systematic literature review. *Journal of Manual & Manipulative Therapy,* 17, 22-35.

HIRAMITSU, T., TOMINAGA, Y., OKADA, M., YAMAMOTO, T. & KOBAYASHI, T. 2015. A Retrospective Study of the Impact of Intraoperative Intact Parathyroid Hormone Monitoring During Total Parathyroidectomy for Secondary Hyperparathyroidism: STARD Study. *Medicine,* 94, e1213.

HOFFMAN, T.C., GLASZIOU, P. P., BOUTRON, I., MILNE, R., PERARA, R., MOHER, D., ALTMAN, D. G., BARBOUR, V., MACDONALD, H., JOHNSTON, M., LAMB, S. E., DIXON-WOODS, M., MCCULLOCH, P., WYATT, J. C. & CHAN, A-W. 2014. Better reporting of interventions: template for intervention description and replication (TIDierR) checklist and guide. *BMJ,* 348:g 1687.

HSU, C.-C. & SANDFORD, B. A. 2007. The Delphi technique: making sense of consensus. *Practical Assessment, Research & Evaluation,* 12, 1-8.

HUDON, C., FORTIN, M., HAGGERTY, J. L., LAMBERT, M. & POITRAS, M.-E. 2011. Measuring patients' perceptions of patient-centered care: a systematic review of tools for family medicine. *The Annals of Family Medicine,* 9, 155-164.

INSTITUTE, S. 2012. *SAS/STAT 12.1 User's Guide Survival Analysis (book Excerpt)*, SAS Institute Incorporated.

JAHROMI, A. S., CINÀ, C. S., LIU, Y. & CLASE, C. M. 2005. Sensitivity and specificity of color duplex ultrasound measurement in the estimation of internal carotid artery stenosis: a systematic review and meta-analysis. *Journal of Vascular Surgery,* 41, 962-972.

JONES, C. M. & ATHANASIOU, T. 2009. Diagnostic accuracy meta-analysis: review of an important tool in radiological research and decision making. *Br J Radiol,* 82, 441-6.

KEMMIS, S. & MCTAGGART, R. 2005. Communicative action and the public sphere. *Denzin, NK & Lincoln, YS (red.), The Sage Handbook of Qualitative Research,* 3, 559-603.

KIM, W. H., CHANG, J. M., MOON, H.-G., YI, A., KOO, H. R., GWEON, H. M. & MOON, W. K. 2015. Comparison of the diagnostic performance of digital breast tomosynthesis and magnetic resonance imaging added to digital mammography in women with known breast cancers. *European Radiology*, 1-9.

KINDON, S. & ELWOOD, S. 2009. Introduction: More than Methods—Reflections on Participatory Action Research in Geographic Teaching, Learning and Research: Participatory Action Research in Geographic Teaching, Learning and Research. *Journal of Geography in Higher Education,* 33, 19-32.

KNOTTNERUS, J. A. & MURIS, J. W. 2003. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *Journal of Clinical Epidemiology,* 56, 1118-1128.

KOH, K.-J., LIST, T., PETERSSON, A. & ROHLIN, M. 2008. Relationship between clinical and magnetic resonance imaging diagnoses and findings in degenerative and inflammatory temporomandibular joint diseases: a systematic literature review. *Journal of Orofacial Pain,* 23, 123-139.

KOREVAAR, D. A., WANG, J., VAN ENST, W. A., LEEFLANG, M. M., HOOFT, L., SMIDT, N. & BOSSUYT, P. M. 2014. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology,* 274, 781-789.

KOSHY, E., KOSHY, V. & WATERMAN, H. 2011. Action Research in Healthcare. London: Sage.

KUNATH, F., GROBE, H. R., RÜCKER, G., ENGEHAUSEN, D., ANTES, G., WULLICH, B. & MEERPOHL, J. J. 2012. Do journals publishing in the field of urology endorse reporting guidelines? A survey of author instructions. *Urologia Internationalis,* 88, 54-59.

LAMÉRIS, W., VAN RANDEN, A., VAN ES, H. W., VAN HEESEWIJK, J. P. M., VAN RAMSHORST, B., BOUMA, W. H., TEN HOVE, W., VAN LEEUWEN, M. S., VAN KEULEN, E. M., DIJKGRAAF, M. G. W., BOSSUYT, P. M. M., BOERMEESTER, M. A. & STOKER, J. 2009. *Imaging strategies for detection of urgent conditions in patients with acute abdominal pain: diagnostic accuracy study*. *BMJ*. 338:b2431.

LANGLOIS, S., GOUDREAU, J. & LALONDE, L. 2014. Scientific rigour and innovations in participatory action research investigating workplace learning in continuing interprofessional education. *Journal of Interprofessional Care,* 28, 226-231.

LAWRENCE, R. J. & DESPRÉS, C. 2004. Futures of transdisciplinarity. *Futures,* 36, 397-405.

LEEFLANG, M. M. 2015. Reporting diagnostic accuracy studies: where are we now? *Biomarkers in Medicine,* 9, 897-899.

LEEFLANG, M. M. G., BOSSUYT, P. M. M. & IRWIG, L. 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of Clinical Epidemiology,* 62, 5-12.

LEES, R., SELVARAJAH, J., FENTON, C., PENDLEBURY, S. T., LANGHORNE, P., STOTT, D. J. & QUINN, T. J. 2014. Test accuracy of cognitive screening tests for diagnosis of dementia and multidomain cognitive impairment in stroke. *Stroke,* 45, 3008-3018.

LENTLE, B. C., ARNOLD, R. E., BECKER, G. J., BRYAN, R. N., FRITZSCHE, P. J. & HUSSEY, D. H. 2007. What We Do Not Yet Know in the Radiologic Sciences 1. *Radiology,* 243, 618-621.

LIGOCKI, C., ABADEH, A., WANG, K. C., ADAMS-WEBBER, T. & DORIA, A. 2015. A systematic review of ultrasound imaging as a tool for evaluating hemophilic arthropathy in children. *Pediatric Radiology,* 45, S234.

LIJMER, J. G., MOL, B. W., HEISTERKAMP, S., BONSEL, G. J., PRINS, M. H., VAN DER MEULEN, J. H. & BOSSUYT, P. M. 1999. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA,* 282, 1061-1066.

LUMBRERAS, B., JARRÍN, I. & HERNÁNDEZ AGUADO, I. 2006. Evaluation of the research methodology in genetic, molecular and proteomic tests. *Gaceta Sanitaria,* 20, 368-373.

LUNDH, A. & GØTZSCHE, P. C. 2008. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC Medical Research Methodology,* 8, 22.

MACKENZIE, N. & KNIPE, S. 2006. Research dilemmas: Paradigms, methods and methodology. *Issues in Educational Research,* 16, 193-205.

MACLEAN, E. N., STONE, I. S., CEELEN, F., GARCIA-ALBENIZ, X., SOMMER, W. H. & PETERSEN, S. E. 2014. Reporting standards in cardiac MRI, CT, and SPECT diagnostic accuracy studies: analysis of the impact of STARD criteria. *European Heart Journal-Cardiovascular Imaging*, 15(6), 691-700.

MAHONEY, J. & ELLISON, J. 2007. Assessing the quality of glucose monitor studies: a critical evaluation of published reports. *Clinical Chemistry,* 53, 1122-1128.

MALCIUS, D., JONKUS, M., KUPRIONIS, G., MALECKAS, A., MONASTYRECKIENE, E., UKTVERIS, R., RINKEVICIUS, S. & BARAUSKAS, V. 2009. The accuracy of different imaging techniques in diagnosis of acute hematogenous osteomyelitis. *Medicina (Kaunas),* 45, 624-631.

MANCHIKANTI, L., DERBY, R., WOLFER, L., SINGH, V., DATTA, S. & HIRSCH, J. A. 2009. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 5. Diagnostic accuracy studies. *Pain Physician,* 12, 517-40.

MARTIN, J., WILLIAMS, K., SUTTON, A. J., ABRAMS, K. R. & ASSASSA, R. 2006. Systematic review and meta-analysis of methods of diagnostic assessment for urinary incontinence. *Neurourology and Urodynamics,* 25, 674-683.

MCCOMISKEY, M. H., MCCLUGGAGE, W. G., GREY, A., HARLEY, I., DOBBS, S. & NAGAR, H. A. 2012. Diagnostic accuracy of magnetic resonance imaging in endometrial cancer. *International Journal of Gynecological Cancer,* 22, 1020-1025.

MCNAMEE, L. S., O'BRIEN, F. Y. & BOTHA, J. H. 2009. Student perceptions of medico-legal autopsy demonstrations in a student-centred curriculum. *Medical Education,* 43, 66-73.

MCNIFF, J. 1995. *Action research for professional development*, Hyde Publications Bournemouth.

MCNIFF, J. 2002. Action research for professional development Concise advice for new action researchers.

MCNIFF, J. 2013. *Action research: Principles and Practice*, Routledge.

MCNIFF, J. & WHITEHEAD, A. 2002. Action Research: Principles and Practice.

MEDINA, L. S. & BLACKMORE, C. C. 2007. Evidence-based Radiology: Review and Dissemination 1. *Radiology,* 244, 331-336.

MEDINA, L. S. & ZURAKOWSKI, D. 2003. Measurement Variability and Confidence Intervals in Medicine: Why Should Radiologists Care?1. *Radiology,* 226, 297-301.

MENON, P., GEEVASINGA, N., YIANNIKAS, C., HOWELLS, J., KIERNAN, M. C. & VUCIC, S. 2015. Sensitivity and specificity of threshold tracking transcranial magnetic stimulation for diagnosis of amyotrophic lateral sclerosis: a prospective study. *The Lancet Neurology,* 14, 478-484.

MEYER, J. 2000. Qualitative research in health care: Using qualitative methods in health related action research. *BMJ: British Medical Journal,* 320, 178.

MIERITZ, R. M., BRONFORT, G., KAWCHUK, G., BREEN, A. & HARTVIGSEN, J. 2012. Reliability and measurement error of 3-dimensional regional lumbar motion measures: a systematic review. *Journal of Manipulative and Physiological Therapeutics,* 35, 645-656.

MILLER, E., ROPOSCH, A., ULERYK, E. & DORIA, A. S. 2009. Juvenile Idiopathic Arthritis of Peripheral Joints: Quality of Reporting of Diagnostic Accuracy of Conventional MRI1. *Academic Radiology,* 16, 739-757.

MITCHELL, A. J. & COYNE, J. C. 2007. Do ultra-short screening instruments accurately detect depression in primary care? *British Journal of General Practice,* 57, 144-151.

MOHER, D., ALTMAN, D., SCHULZ, K., SIMERA, I. & WAGER, E. 2014a. *Guidelines for Reporting Health Research: A User's Manual*, Wiley Online Library.

MOHER, D., ALTMAN, D. G., SCHULZ, K. F. & SIMERA, I. 2014b. How to Develop a Reporting Guideline. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

MOHER, D., PLINT, A. C., ALTMAN, D. G., SCHULZ, K. F., KOBER, T., GALLOWAY, E. K., WEEKS, L. & DIAS, S. 2010a. Consolidated standards of reporting trials (CONSORT) and the quality of reporting of randomized controlled trials. *The Cochrane Library*.

MOHER, D., SCHULZ, K. F., ALTMAN, D. G., HOEY, J., GRIMSHAW, J., MILLER, D., SEELY, D., SIMERA, I., SAMPSON, M., WEEKS, L. & OCAMPO, M. 2014c. Characteristics of Available Reporting Guidelines. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

MOHER, D., SCHULZ, K. F., SIMERA, I. & ALTMAN, D. G. 2010b. Guidance for developers of health research reporting guidelines. *PLoS medicine,* 7, e1000217.

MOHER, D., WEEKS, L., OCAMPO, M., SEELY, D., SAMPSON, M., ALTMAN, D. G., SCHULZ, K. F., MILLER, D., SIMERA, I. & GRIMSHAW, J. 2011. Describing reporting guidelines for health research: a systematic review. *Journal of Clinical Epidemiology,* 64, 718-742.

MONTORI, V. M., JAESCHKE, R., SCHÜNEMANN, H. J., BHANDARI, M., BROZEK, J. L., DEVEREAUX, P. J. & GUYATT, G. H. 2004. Users' Guide To Detecting Misleading Claims In Clinical Research Reports. *BMJ: British Medical Journal,* 329, 1093-1096.

MORGAN, D. L. 2007. Paradigms lost and pragmatism regained methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research,* 1, 48-76.

MORRISON, B. & LILFORD, R. 2001. How can action research apply to health services? *Qualitative Health Research,* 11, 436-449.

MYBURGH, C., LARSEN, A. H. & HARTVIGSEN, J. 2008. A systematic, critical review of manual palpation for identifying myofascial trigger points: evidence and clinical significance. *Archives of Physical Medicine and Rehabilitation,* 89, 1169-1176.

NAGAR, H., DOBBS, S., MCCLELLAND, H. R., PRICE, J., MCCLUGGAGE, W. G. & GREY, A. 2006. The diagnostic accuracy of magnetic resonance imaging in detecting cervical involvement in endometrial cancer. *Gynecologic Oncology,* 103, 431-434.

NETWORK, E. 2009. Enhancing the quality and transparency of health research. *Available at www. equator-network.org*

NIEUWENHUIJZE, M. J., KORSTJENS, I., DE JONGE, A., DE VRIES, R. & LAGRO-JANSSEN, A. 2014. On speaking terms: a Delphi study on shared decision-making in maternity care. *BMC Pregnancy and Childbirth,* 14, 223-223.

NOEL-STORR, A. H., FLICKER, L., RITCHIE, C. W., NGUYEN, G. H., GUPTA, T., WOOD, P., WALTON, J., DESAI, M., SOLOMON, D. F. & MOLENA, E. 2013. Systematic review of the body of evidence for the use of biomarkers in the diagnosis of dementia. *Alzheimer's & Dementia,* 9, e96-e105.

O'LEARY, J. D. & CRAWFORD, M. W. 2013. Review article: Reporting guidelines in the biomedical literature. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie,* 60, 813-821.

OBUCHOWSKI, N. A. 2003a. Receiver operating characteristic curves and their use in radiology. *Radiology,* 229, 3-8.

OBUCHOWSKI, N. A. 2003b. Special Topics III: Bias. *Radiology,* 229, 617-621.

OCHODO, E. A., DE HAAN, M. C., REITSMA, J. B., HOOFT, L., BOSSUYT, P. M. & LEEFLANG, M. M. 2013. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology,* 267, 581-8.

PARK, P. 2006. Knowledge and Participatory Research. *Handbook of Action Research: Concise paperback edition*, 83-93.

PARRY, R. A., GLAZE, S. A. & ARCHER, B. R. 1999. The AAPM/RSNA physics tutorial for residents. Typical patient radiation doses in diagnostic radiology. *Radiographics : a review publication of the Radiological Society of North America, Inc,* 19, 1289.

PAVLOV, C., CASAZZA, G., NIKOLOVA, D., TSOCHATZIS, E., BURROUGHS, A., IVASHKIN, V. & GLUUD, C. 2015. Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease. *Cochrane Database Syst Rev*. 22;1:CD010542.

PERRY, A. E., MARANDOS, R., COULTON, S. & JOHNSON, M. 2010. Screening tools assessing risk of suicide and self-harm in adult offenders: a systematic review. *International Journal of Offender Therapy and Comparative Criminology*.

PLOUS, S. 1993. *The psychology of judgment and decision making*, Mcgraw-Hill Book Company.

POLDRACK, R. A., FLETCHER, P. C., HENSON, R. N., WORSLEY, K. J., BRETT, M. & NICHOLS, T. E. 2008. Guidelines for reporting an fMRI study. *NeuroImage,* 40, 409-414.

POPE, A. 2009. Reproducibility: Intraobserver and Interobserver Variability. Biostatistics for Radiologists, 125-140.

PRUMMEL, M. V., MURADALI, D., SHUMAK, R., MAJPRUZ, V., BROWN, P., JIANG, H., DONE, S. J., YAFFE, M. J. & CHIARELLI, A. M. 2015. Digital Compared with Screen-Film Mammography: Measures of Diagnostic Accuracy among Women Screened in the Ontario Breast Screening Program. *Radiology*, 150733.

RAHMAN, R. L., CRAWFORD, S. L. & SIWAWA, P. 2015. Management of axilla in breast cancer–The saga continues. *The Breast*.

RAJA, A. S., PINES, J. M., SCHUUR, J. D., MUIR, M., CALFEE, R. P. & CARPENTER, C. R. 2013. Evidence based diagnostics: Meta-analysis of the accuracy of physical exam and imaging for adult scaphoid fractures. *Academic Emergency Medicine,* 20, S24-S25.

RAMA, K. R. B. S., POOVALI, S. & APSINGI, S. 2006. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clinical Orthopaedics and Related Research,* 447, 237-246.

RAUTIAINEN, S., MASARWAH, A., SUDAH, M., SUTELA, A., PELKONEN, O., JOUKAINEN, S., SIRONEN, R., KÄRJÄ, V. & VANNINEN, R. 2013. Axillary lymph node biopsy in newly diagnosed invasive breast cancer: comparative accuracy of fine-needle aspiration biopsy versus core-needle biopsy. *Radiology,* 269, 54-60.

RCR. 2012. *Good practice guide for clinical radiologists* [Online]. London. Available: https://www.rcr.ac.uk/sites/default/files/publication/BFCR%2812%291_GoodPractice.pdf.

REASON, P. & BRADBURY, H. 2001. Introduction: Inquiry and Participation in Search of a World Worthy of Human Aspiration, w: P. Reason, H. Bradbury (red.). *Handbook of Action Research Participative Inquiry and Practice*.

REASON, P. & BRADBURY, H. 2005. *Handbook of Action Research: concise paperback edition*, Sage.

REITSMA, J. B., RUTJES, A. W. S., KHAN, K. S., COOMARASAMY, A. & BOSSUYT, P. M. 2009. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology,* 62, 797-806.

RENNIE, D. 2014. Frontmatter. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

RIEGELMAN, R. K. 2000. *Studying a study and testing a test: how to read the medical evidence,* Philadelphia, Lippincott Williams & Wilkins.

RIEGELMAN, R. K. 2013. Testing a Test - M.A.A.R.I.E. Framework: Method, Assignment, and Assessment *In:* RHYNER, S. (ed.) *Studying A Study & Testing A Test.* Baltimore, MD Lippincott Williams & Wilkins

ROBSON, C. 2002. *Real world research: A resource for social scientists and practitioner-researchers*, Blackwell Oxford.

ROPOSCH, A., MOREAU, N. M., ULERYK, E. & DORIA, A. S. 2006. Developmental dysplasia of the hip: quality of reporting of diagnostic accuracy for US. *Radiology,* 241, 854-860.

RUTJES, A. W. S., REITSMA, J. B., DI NISIO, M., SMIDT, N., VAN RIJN, J. C. & BOSSUYT, P. M. M. 2006. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne,* 174, 469-476.

RYCHETNIK, L., FROMMER, M., HAWE, P. & SHIELL, A. 2002. Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health,* 56, 119-127.

SABA, L., GUERRIERO, S., SULCIS, R., PILLONI, M., AJOSSA, S., MELIS, G. & MALLARINI, G. 2012. MRI and “Tenderness Guided” transvaginal ultrasonography in the diagnosis of recto-sigmoid endometriosis. *Journal of Magnetic Resonance Imaging,* 35, 352-360.

SABA, L., GUERRIERO, S., SULIS, R., PILLONI, M., AJOSSA, S., MELIS, G. & MALLARINI, G. 2011. Learning curve in the detection of ovarian and deep endometriosis by using Magnetic Resonance: comparison with surgical results. *European Journal of Radiology,* 79, 237-244.

SANDREY, M. A. 2013. Special physical examination tests for superior labrum anterior-posterior shoulder tears: an examination of clinical usefulness. *Journal of Athletic Training,* 48, 856.

SARDANELLI, F. & DI LEO, G. 2009a. *Biostatistics for radiologists: planning, performing, and writing a radiologic study*, Springer Science & Business Media.

SARDANELLI, F. & DI LEO, G. 2009b. Reproducibility: Intraobserver and Interobserver Variability. *Biostatistics for Radiologists.* Springer Milan.

SARDANELLI, F., HUNINK, M. G., GILBERT, F. J., DI LEO, G. & KRESTIN, G. P. 2010. Evidence-based radiology: why and how? *European Radiology,* 20, 1-15.

SCHATZKI, T. R., KNORR-CETINA, K. D. & SAVIGNY, E. V. (eds.) 2000. *The Practice Turn in Contemporary Theory*: Taylor & Francis Ltd - M.U.A.

SCHÖN, D. 1987a. *Educating the Reflective Practitioner,* San Francisco, Jossey-Bass.

SCHÖN, D. 1987b. Educating the Reflective Practitioner.

SCHUBERT, T., TAKES, M., ASCHWANDEN, M., KLARHOEFER, M., HAAS, T., JACOB, A. L., LIU, D., GUTZEIT, A. & KOS, S. 2015. Non-enhanced, ECG-gated MR angiography of the pedal vasculature: comparison with contrast-enhanced MR angiography and digital subtraction angiography in peripheral arterial occlusive disease. *European Radiology*, 1-9.

SCOTT, D., BROWN, A. & LUNT, I. 2004. *Professional Doctorates: Integrating Academic And Professional Knowledge: Integrating Academic and Professional Knowledge*, McGraw-Hill Education (UK).

SEITH, F., GATIDIS, S., SCHMIDT, H., BEZRUKOV, I., LA FOUGÈRE, C., NIKOLAOU, K., PFANNENBERG, C. & SCHWENZER, N. 2015. Comparison of Positron Emission Tomography Quantification Using Magnetic Resonance-and Computed Tomography-Based Attenuation Correction in Physiological Tissues and Lesions: A Whole-Body Positron Emission Tomography/Magnetic Resonance Study in 66 Patients. *Investigative Radiology*.

SELMAN, T., KHAN, K. & MANN, C. 2005. An evidence-based approach to test accuracy studies in gynecologic oncology: the ‘STARD’checklist. *Gynecologic Oncology,* 96, 575-578.

SELMAN, T. J., MORRIS, R. K., ZAMORA, J. & KHAN, K. S. 2011. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. *BMC Women's Health,* 11, 8.

SHARMA, S., CROW, H. C., MCCALL JR, W. & GONZALEZ, Y. M. 2012. Systematic review of reliability and diagnostic validity of joint vibration analysis for diagnosis of temporomandibular disorders. *Journal of Orofacial Pain,* 27, 51-60.

SHIVKUMAR, S., PEELING, R., JAFARI, Y., JOSEPH, L. & PAI, N. P. 2012. Accuracy of rapid and point-of-care screening tests for hepatitis C: a systematic review and meta-analysis. *Annals of Internal Medicine,* 157, 558-566.

SHUNMUGAM, M. & AZUARA-BLANCO, A. 2006. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Investigative Ophthalmology & Visual Science,* 47, 2317-2323.

SIDDIQUI, I. A., SABAH, S. A., SATCHITHANANDA, K., LIM, A. K., CRO, S., HENCKEL, J., SKINNER, J. A. & HART, A. J. 2014. A comparison of the diagnostic accuracy of MARS MRI and ultrasound of the painful metal-on-metal hip arthroplasty. *Acta Orthopaedica,* 85, 375-382.

SIDDIQUI, M., AZUARA-BLANCO, A. & BURR, J. 2005. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *British Journal of Ophthalmology,* 89, 261-265.

SIMERA, I., ALTMAN, D. G., MOHER, D., SCHULZ, K. F. & HOEY, J. 2008. Guidelines for reporting health research: the EQUATOR network's survey of guideline authors. *PLoS Med,* 5, e139.

SMIDT, N., RUTJES, A.W., VAN DER WINDT, D. A., OSTELO, R. W., REITSMA, J. B., BOSSUYT, P. M., BOUTER, L. M. & DE VET, H. C. 2005. Quality of Reporting of Diagnostic Accuracy Studies. *Radiology,* 235, 347-353.

SMIDT, N., RUTJES, A., VAN DER WINDT, D., OSTELO, R., BOSSUYT, P., REITSMA, J., BOUTER, L. & DE VET, H. 2006. The quality of diagnostic accuracy studies since the STARD statement: Has it improved? *Neurology,* 67, 792-797.

SMITH, L., BRATINI, L., CHAMBERS, D.-A., JENSEN, R. V. & ROMERO, L. 2010. Between idealism and reality: Meeting the challenges of participatory action research. *Action Research,* 8, 407-425.

SMITH, M. K. 2001. Chris Argyris: theories of action, double-loop learning and organizational learning. *The encyclopedia of informal education,* 1.

SOLOMON, M. 2011. Just a paradigm: evidence-based medicine in epistemological context. *European Journal for Philosophy of Science,* 1, 451-466.

SOMEKH, B. 1995. The contribution of action research to development in social endeavours: a position paper on action research methodology. *British Educational Research Journal,* 21, 339-355.

SOUSA, V. E. C., LOPES, M. V. D. O. & SILVA, V. M. 2015. Systematic review and meta-analysis of the accuracy of clinical indicators for ineffective airway clearance. *Journal of Advanced Nursing,* 71, 498-513.

STENGEL, D., BAUWENS, K., RADEMACHER, G., MUTZE, S. & EKKERNKAMP, A. 2005. Association between Compliance with Methodological Standards of Diagnostic Research and Reported Test Accuracy: Meta-Analysis of Focused Assessment of US for Trauma 1. *Radiology,* 236, 102-111.

STEVENS, A., SHAMSEER, L., WEINSTEIN, E., YAZDI, F., TURNER, L., THIELMAN, J., ALTMAN, D. G., HIRST, A., HOEY, J. & PALEPU, A. 2014. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ: British Medical Journal,* 348.

STOKOLS, D. 2006. Toward a science of transdisciplinary action research. *American Journal of Community Psychology,* 38, 63-77.

STRASSLE, P., HESS, A. S., THOM, K. A. & HARRIS, A. D. 2012. Assessing Sensitivity and Specificity in New Diagnostic Tests: The Importance and Challenges of Study Populations. *Infection Control and Hospital Epidemiology,* 33, 1177-1178.

STREINER, D. L. 2003. Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment,* 81, 209-219.

SU, S.-B., QIN, S.-Y., CHEN, W., LUO, W. & JIANG, H.-X. 2015. Carbohydrate antigen 19-9 for differential diagnosis of pancreatic carcinoma and chronic pancreatitis. *World Journal of Gastroenterology: WJG,* 21, 4323.

SU, S.-B., QIN, S.-Y., GUO, X.-Y., LUO, W. & JIANG, H.-X. 2013. Assessment by meta-analysis of interferon-gamma for the diagnosis of tuberculous peritonitis. *World Journal of Gastroenterology: WJG,* 19, 1645.

THAWATCHAI LEELAHANAJ, M. 2010. Developing Thai economic model to study cost-effectiveness of switching to bupropion compared to combination with bupropion after the failure of an SSRI for major depressive disorder. *J Med Assoc Thai,* 93, S35-S42.

THIEME, M. E., LEEUWENBURGH, M. M., VALDEHUEZA, Z. D., BOUMAN, D. E., DE BRUIN, I. G., SCHREURS, W. H., HOUDIJK, A. P., STOKER, J. & WIARDA, B. M. 2014. Diagnostic accuracy and patient acceptance of MRI in children with suspected appendicitis. *European Radiology,* 24, 630-637.

THORNTON, G., MCPHAIL, M., NAYAGAM, S., HEWITT, M., VLAVIANOS, P. & MONAHAN, K. 2013. Endoscopic ultrasound guided fine needle aspiration for the diagnosis of pancreatic cystic neoplasms: a meta-analysis. *Pancreatology,* 13, 48-57.

TSANG, A. C., PIRSHAHID, S. A., VIRGILI, G., GOTTLIEB, C. C., HAMILTON, J. & COUPLAND, S. G. 2015. Hydroxychloroquine and Chloroquine Retinopathy: A Systematic Review Evaluating the Multifocal Electroretinogram as a Screening Test. *Ophthalmology,* 122, 1239-1251. e4.

TSENG, D. S., VAN SANTVOORT, H. C., OFFERHAUS, G. J. A., BESSELINK, M. G., BOREL RINKES, I. H., VAN LEEUWEN, M. S. & MOLENAAR, I. Q. 2015. The role of CT in assessment of extra-regional lymph node involvement in pancreatic and peri-ampullary cancer: A prospective diagnostic accuracy study. *Pancreatology,* 15, S85-S86.

UIJL, S. G., LEIJTEN, F. S., PARRA, J., ARENDS, J. B., VAN HUFFELEN, A. C. & MOONS, K. G. 2005. What is the current evidence on decision-making after referral for temporal lobe epilepsy surgery?: A review of the literature. *Seizure,* 14, 534-540.

VAN DEN BRUEL, A., AERTGEERTS, B. & BUNTINX, F. 2006. Results of diagnostic accuracy studies are not always validated. *Journal of Clinical Epidemiology,* 59, 559.e1-559.e9.

VAN ERKEL, A. R. & PETER, M. 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of Radiology,* 27, 88-94.

VAN TRIJFFEL, E., ANDEREGG, Q., BOSSUYT, P. & LUCAS, C. 2005. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual Therapy,* 10, 256-269.

VILLAR, C. G. 2011. Evidence-based radiology for diagnostic imaging: What it is and how to practice it. *Radiología (English Edition),* 53, 326-334.

WALTHER, S., SCHUELER, S., TACKMANN, R., SCHUETZ, G. M., SCHLATTMANN, P. & DEWEY, M. 2014. Compliance with STARD checklist among studies of coronary CT angiography: systematic review. *Radiology,* 271, 74.

WANG, K., MUNIR, S., SHARIATI, K., ADAMS-WEBBER, T. & DORIA, A. 2015. Clinical utility of dual-energy x-ray absorptiometry for assessment of fractures in pediatric osteoporosis: Evidence-based knowledge synthesis. *Pediatric Radiology,* 45, S219-S221.

WANG, W., CHEN, L.-D., LU, M.-D., LIU, G.-J., SHEN, S.-L., XU, Z.-F., XIE, X.-Y., WANG, Y. & ZHOU, L.-Y. 2013. Contrast-enhanced ultrasound features of histologically proven focal nodular hyperplasia: diagnostic performance compared with contrast-enhanced CT. *European Radiology,* 23, 2546-2554.

WARDLAW, J., CHAPPELL, F., BEST, J., WARTOLOWSKA, K. & BERRY, E. 2006. Non-invasive imaging compared with intra-arterial angiography in the diagnosis of symptomatic carotid stenosis: a meta-analysis. *The Lancet,* 367, 1503-1512.

WARDLAW, J. M., BRINDLE, W., CASADO, A. M., SHULER, K., HENDERSON, M., THOMAS, B., MACFARLANE, J., MANIEGA, S. M., LYMER, K. & MORRIS, Z. 2012. A systematic review of the utility of 1.5 versus 3 Tesla magnetic resonance brain imaging in clinical practice and research. *European Radiology,* 22, 2295-2303.

WATERMAN, H., TILLEN, D., DICKSON, R. & DE KONING, K. 2000. Action research: a systematic review and guidance for assessment. *Health Technology Assessment (Winchester, England),* 5, iii-157.

WHITING, P., RUTJES, A. W., REITSMA, J. B., BOSSUYT, P. M. & KLEIJNEN, J. 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol,* 3, 25.

WIDDIFIELD, J., LABRECQUE, J. & LIX, L. A Systematic Review to Evaluate the Quality and Reporting of Administrative Database Validation Studies for Rheumatic Diseases. 2011. *JOURNAL OF RHEUMATOLOGY,* 2011.1177-1178.

WIESKE, L., VERHAMME C, INHENFELDT, D., VAN DER SCHAAF, M., BOUWES, A. & SCHULTZ, M. 2012. Prediction of intensive care unit-aquired weakness using a simplified electrophysiological screening test. *Am J Respir Crit Care*, 185.

WILCZYNSKI, N. L. 2008. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. *Radiology,* 248, 817-823.

WILCZYNSKI, N. L. & HAYNES, R. B. 2007a. Indexing of Diagnostic Accuracy Studies in MEDLINE and EMBASE. *AMIA Annual Symposium Proceedings,* 2007, 801-805.

WILCZYNSKI, N. L. & HAYNES, R. B. Indexing of Diagnostic Accuracy Studies in MEDLINE and EMBASE. AMIA Annual Symposium Proceedings, 2007b. American Medical Informatics Association, 801.

WILCZYNSKI, N. L., MCKIBBON, K. A., WALTER, S. D., GARG, A. X. & HAYNES, R. B. 2013a. MEDLINE clinical queries are robust when searching in recent publishing years. *Journal of the American Medical Informatics Association : JAMIA,* 20, 363-368.

WILCZYNSKI, N. L., MCKIBBON, K. A., WALTER, S. D., GARG, A. X. & HAYNES, R. B. 2013b. MEDLINE clinical queries are robust when searching in recent publishing years. *Journal of the American Medical Informatics Association,* 20, 363-368.

WU, L., LI, Y., LI, Z., CAO, Y. & GAO, F. 2013. Diagnostic accuracy of narrow-band imaging for the differentiation of neoplastic from non-neoplastic colorectal polyps: a meta-analysis. *Colorectal Disease,* 15, 3-11.

YANG, D. H., KIM, Y.-H., ROH, J.-H., KANG, J.-W., HAN, D., JUNG, J., KIM, N., LEE, J. B., AHN, J.-M. & LEE, J.-Y. 2015. Stress Myocardial Perfusion CT in Patients Suspected of Having Coronary Artery Disease: Visual and Quantitative Analysis—Validation by Using Fractional Flow Reserve. *Radiology*, 141126.

ZAFAR, A., KHAN, G. I. & SIDDIQUI, M. 2008. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. *Clinical & Experimental Ophthalmology,* 36, 537-542.

ZHANG, L. J., WANG, Y., SCHOEPF, U. J., MEINEL, F. G., BAYER II, R. R., QI, L., CAO, J., ZHOU, C. S., ZHAO, Y. E. & LI, X. 2015. Image quality, radiation dose, and diagnostic accuracy of prospectively ECG-triggered high-pitch coronary CT angiography at 70 kVp in a clinical setting: comparison with invasive coronary angiography. *European Radiology*, 1-10.

Appendix 1: Library Search September 1, 2013 to October 31, 2013

Library search that was used for the four high-impact radiology journals from September 1, 2013 to October 31, 2013 which resulted in 71 citations:

((((((di[MeSH Subheading]) OR diagnos*) OR (("Sensitivity and Specificity"[MeSH Terms]))) OR stard) OR Standards for Reporting of Diagnostic Accuracy)) AND ((((("Radiology"[Journal]) OR "Investigative radiology"[Journal]) OR "The British journal of radiology"[Journal]) OR "European radiology"[Journal]) AND ("2013/09/01"[PDat] : "2013/10/31"[PDat]))

Appendix 2: Library Search September 1, 2015 to October 31, 2015

Library search that was used for the same four journals from September 1, 2015 to October 31, 2015, which resulted in 326 citations:

((((((di[MeSH Subheading]) OR (diagnos*)) OR (("Sensitivity and Specificity"[MeSH Terms]))) OR stard) OR Standards for Reporting of Diagnostic Accuracy)) AND ((((("Radiology"[Journal]) OR "Investigative radiology"[Journal]) OR "The British journal of radiology"[Journal]) OR "European radiology"[Journal] AND (("2015/09/01"[PDat] : "2015/10/31"[PDat])) = 326

Appendix 3: RadSTARD 2 Page Summary Document

RadSTARD –Radiology Standards for the Reporting of Diagnostic Accuracy Studies Summary Document

1. TITLE & ABSTRACT	Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract. *Use the term ‘diagnostic accuracy’ in the title of the study (STARD item # 1).*
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare the index test with the reference standard for diagnosis of a specific condition.** *The aim is to compare the index test to the reference standard (STARD item # 2).*
3. METHODS, PATIENT ELIGIBILITY, DATA COLLECTION	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided. *Specify whether patients studied were receiving primary or secondary care (STARD item # 3).*
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.** *Specify if the participants received either the index test or the reference standard (STARD item # 3).*
5.	Was patient selection consecutive or non-consecutive? *Specify how participants were enrolled (STARD item # 5).*
6.	**Data collection (prospective or retrospective) should be provided with start dates and end dates.** *Record that data was collected as a chart review (retrospectively) or prospective analysis and include the dates the study was done (STARD item # 6 and 14).*
7. SAMPLE SIZE	**Sample size and limitations should be provided.** *Providing a sample size reflects the power of the study results (not in the STARD).*

8. DIAGNOSTIC TEST METHODS	Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale. *Reporting the reference standard which is the test that is commonly used in medical practice. Rationale for its use should also be defined (STARD item # 7).*
9. IMPERFECT REFERENCE STANDARD	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified.** *There is no perfect reference standard – alternatives may consist of a panel standard, composite reference standard, latent class analysis or intrinsic reference standard (not in STARD).*
10. TEST	Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies. *A thorough description of the index test and reference standard should be provided so that the reader can interpret if the same test could be performed in their institution (STARD item # 8).*
11.	**Sound theoretical physics basis of the index test should be provided for new techniques.** *Description of physics for the index test is very pertinent to radiology to facilitate replication (not in STARD).*
12.	**Modifications during the study should be reported if they occurred.** *Describe if the index test or reference standard was modified as this could impact diagnostic accuracy and alter data analysis (not in STARD).*
13. ANALYSIS	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests. *It is important to specify which cut-off values were used for a specific diagnosis (similar to STARD item # 9).*
14.	The training and number of investigators should be described including **any extra training for new techniques.** *Whether any extra training was required to interpret the index test should be known (similar to STARD item # 10).*
15.	Whether readers were blinded to prior test results or clinical information should be known. *The blinding of interpreters is essential to prevent bias (STARD item # 11).*
16. RESULTS	Radiology Diagnostic Accuracy Trials should report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations. *The provision of a flow diagram clearly illustrates how many participants were tested or not resulting in final analysis (STARD item # 16).*
17.	Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and **concurrent therapies** should be provided. *The provision of concurrent therapies should be provided as they may affect the interpretation of the tests (similar to STARD item # 15).*
18.	The severity (spectrum) of the disease entity should be explicitly reported. *The time that a disease was present or if was found by screening could impact how readily it is diagnosed (STARD item # 18).*
19.	The time difference between the index test and reference test and details of any other treatments provided between the two tests should be provided. *If there was a delay between the conduct of either test this could impact their level of diagnostic accuracy (STARD item # 17).*
20.	Adverse events should be reported for either the index test or reference standard. *It is important to state if there were any incidents of adverse events when conducting either test (STARD item # 20).*
21. STATISTICS	Radiology Diagnostic Accuracy Trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI). *Calculated confidence intervals and p-values define how precise the estimates are for the population chosen under study (similar to STARD item # 21).*
22.	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers for the index test **and reference standard**; and describe how this data was handled. *All indeterminate results from the index test and the reference standard should be reported as ignoring such results can bias measures of diagnostic accuracy (similar to STARD item # 22).*
23.	Radiology Diagnostic Accuracy Trials should include estimates of **inter-observer** variability **including > 3 observers with variable expertise.** *A description of analysis that is planned should be described apriori with > 3 observers of variable expertise as this will decrease inter-observer variability (similar to STARD item # 23).*
24.	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility. *This is done by stating the level of Kappa where the statistic impacts the level of intra and inter-observer variability (STARD item # 24).*
25. DISCUSSION	The clinical relevance of the study findings should be provided. *Discussion could include comparing and contrasting to previous studies plus describing any limitations to the current study (STARD item # 25).*

Appendix 4: RadSTARD Elaboration Document

The RadSTARD – Radiology Standards for Reporting of Diagnostic Accuracy Studies Explanation and Elaboration Document

1. TITLE & ABSTRACT	**Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract.**
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare the index test with the reference standard for diagnosis of a specific condition.**
3. METHODS **PATIENT ELIGIBILITY, DATA COLLECTION**	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided.
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.**
5.	Was patient selection consecutive or non-consecutive?
6.	**Data collection (prospective or retrospective) should be provided with start dates and end dates.**
7. SAMPLE SIZE	**Sample size and limitations should be provided.**
8. DIAGNOSTIC TEST METHODS	Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale.
9. IMPERFECT REFERENCE STANDARD	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified.**
10. TEST	Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies.
11.	**Sound theoretical physics basis of the index test should be provided for new techniques.**
12.	**Modifications during the study should be reported if they occurred.**
13. ANALYSIS	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests.
14.	The training and number of investigators should be described including **any extra training for new techniques.**
15.	Whether readers were blinded to prior test results or clinical information should be known.
16. RESULTS	Radiology Diagnostic Accuracy Trials should report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations.
17.	Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and **concurrent therapies** should be provided.
18.	The severity (spectrum) of the disease entity should be explicitly reported.

19.	The time difference between the index test and reference test and details of any other treatments provided between the two tests should be provided.
20.	Adverse events should be reported for either the index test or reference standard.
21. STATISTICS	Radiology Diagnostic Accuracy Trials .should explicitly state measures of diagnostic accuracy and uncertainty (preferably **p-values** and 95% CI).
22.	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers for the index test **and reference standard**; and describe how this data was handled.
23.	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability **including > 3 observers with variable expertise.**
24.	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.
25. DISCUSSION	The clinical relevance of the study findings should be provided.

Bossuyt (2003)

Item 1 of the RadSTARD is different. Items 2, 3, 4, 13, 14, 17, 21, 22 and 23 are partially different to the STARD as per bolded insert.

Introduction

Within the discipline of medicine the premise that clinical practice should be based on critical evaluation of published literature has been established for several decades. Applying evidence from clinical studies into clinical practice is commonly referred to as evidence based medicine. Although evidence-based medicine has entered into the discipline of radiology, a relative delay has been noted. This delay may be due to the fact that comparing the diagnostic accuracy of two imaging modalities is far different than comparing the efficacy of a new drug versus a placebo or standard of care. Determining the accuracy of a diagnostic imaging test requires studying the test in a cohort of patients with the suspected medical condition under study as opposed to conducting a randomized controlled trial. This so called practice has been coined evidence-based radiology as it is specific to radiologists. As radiologists interpret imaging, a full comprehension of the implications of their findings in context to the evidence reported in the literature is required (Sardanelli et al., 2010).

Diagnostic Accuracy Studies

In diagnostic accuracy studies the results of the index test are compared to the results obtained from the definitive gold standard or reference standard (Riegelman, 2013). Accuracy is determined by the level of agreement that is bound by the index test or test under evaluation as it is compared to the reference standard. Therefore; diagnostic accuracy can be illustrated in a number of ways including sensitivity and specificity, ROC (receiver operator characteristics) curves, likelihood or odds ratios. Diagnostic accuracy studies compare the results of the index test to the reference standard that are performed on the same individual within a consecutive series of patients (Manchikanti et al., 2009). The results of diagnostic accuracy studies are considered vital in the assessment of new or existing diagnostic tools. As the results of diagnostic accuracy studies are often used to guide future patient care it is paramount that the quality of reporting for the trial results is done completely and accurately. The benefit of accurately reporting trial results are twofold in that it enables the reader to assess for potential bias while simultaneously evaluating the overall generalizability of the study results (Bossuyt et al., 2003a).

Minimizing Bias

When radiologic studies are reviewed by radiology researchers, journal editors or readers of journals they should be aware that the literature they are reading may be reported within inherent biases. Therefore it's important for all who are involved in radiological research to the aware of the potential biases and to learn how to avoid biases from occurring. Bias is noted when the measurements provided for the sensitivity and specificity of a test do not correspond with the values that one would expect if the same test was performed in a similar patient population. As not all patients in a given population are exposed to the test it is pertinent that studies are designed with the aim of mitigating bias (Obuchowski, 2003b).

Throughout the conduct of radiologic study such as diagnostic accuracy trials the incidence of bias can occur on many levels. For instance, patients selected to participate in the study, radiological interpreters chosen to interpret the tests, choosing the reference standard and subsequent analysis of the test results. It is therefore recommended that researchers are aware of the types of biases that could occur in the planning stages of their study and try to avoid them (Obuchowski, 2003b). Many of these biases will be discussed within the RadSTARD with suggestions provided on how to avoid them.

STARD – Standards for Reporting of Diagnostic Accuracy

The STARD statement is a reporting tool that was developed in 2003 by a group of scientists for clinicians to systematically review the results of diagnostic accuracy trials. The tool consists of a list of 25 items plus a flow diagram for authors to utilize to ensure all of the pertinent study information is accurately reported (Bossuyt et al., 2003a).

As diagnostic accuracy is not comprised of a compilation of predetermined examinations, clinicians need to know which test should be used to accurately provide the best treatment for a given condition. If diagnostic accuracy studies are poorly reported this inhibits the objective appraisal by the reader and overall generalizability of the study findings. In addition, if only the favourable aspects of the study findings are presented this may encourage premature early adoption of tests that should not be considered as first-line intervention resulting in delayed diagnosis and wasting valuable healthcare resources.

Numerous reasons have been postulated why there's been such a slow adoption of the STARD statement. It takes time for new guidelines to be adopted by journals and authors. In addition it has been found that journals have given poor instructions to authors on how to incorporate the use of the STARD checklist. Suffice to say there is a gap in the literature in that recommendations are necessary to improve adherence to the guidelines when reporting diagnostic accuracy findings. As described by Ochodo et al. (2013) guidelines such as the STARD are not static and therefore it's possible that amendments to the guidelines would be beneficial to various subspecialties. Although the CONSORT checklist has been amended twice since its initial publication the STARD checklist has remained the same.

Arguably, the adoption of the STARD tool for use when reporting diagnostic accuracy studies specific to radiology has been slowly adopted as reported by those who have conducted systematic reviews assessing its use (Wilczynski, 2008). Concomitantly, other systematic reviews have been reported in the literature whereby a modified checklist of the STARD was utilized to conduct their reviews (Walther et al., 2014, Wilczynski, 2008).

Radiology Diagnostic Accuracy Studies

As the Royal College of Radiology (UK) recommends the use of published clinical trials as a guide in achieving standards of accuracy it is important for radiologists to reflect upon the results from diagnostic accuracy studies (RCR, 2012). This will decrease the use of inappropriate investigations resulting in enhanced patient care as the proper diagnostic test was chosen based on the results from accurate reporting of sensitivity and specificity analysis of diagnostic accuracy trials. Within healthcare the purpose of testing is to provide a constellation of information that is necessary to the decision-making processes. In order to comprehend the results of each test it is essential to understand the multiple purposes a single test can provide. The same test can be used in numerous situations depending on the purpose of why it was ordered. Regardless of which test is chosen the term index test refers to the test results that are being evaluated (Riegelman, 2013). As diagnostic radiology is influenced by the vast amount of evolving technologies within the radiological literature it is important that the results of the findings are accurately interpreted allowing for use in clinical application (Jones and Athanasiou, 2009).

Statistical findings are commonly represented in radiology journals and other evidence-based medical literature. Unfortunately the statistical measures can have errors which many investigators including radiologists are not well prepared to interpret statistical flaws. This could mainly be due to the fact that research methodology is a low priority when training radiologists. In addition, there is also an indifference towards teaching statistical methods in medical school during a physicians training (Applegate and Crewson, 2002).

The Use of STARD in Radiology Diagnostic Accuracy Trials

The use of STARD in the reporting of diagnostic accuracy trials in radiology has been steadily on the rise however; the adoption of this tool has been less than favourable. Concomitantly, it is important to point out that within the literature researchers have been criticized for over interpreting their trial results – the so called ‘evidence of spin’ (Ochodo et al., 2013). Therefore; it is recommended that systematic tools such as the STARD checklist are followed when reporting the study findings so that the results are not misrepresented and that clinicians can make treatment decisions with confidence based on the study findings.

The use of a checklist such as the STARD can assist researchers in the following tri-partate fashion:

The checklist provides the researcher with the important elements of study design to be cognizant of when considering conducting a diagnostic accuracy study.

It illustrates which key items should be emphasized throughout the conduct of the study.

Plus, it encourages communicating study findings more efficiently and accurately (Gatsonis, 2003).

Encouraging the use of checklists as a guide when developing a research protocol enhances the requirements of the essential elements to report later when the study is completed. This in turn will enhance the quality of reporting diagnostic accuracy study findings. Studies that are well reported will also provide information that is reported accurately guiding clinicians in their treatment choices in managing their patients care (Rutjes et al., 2006).

The RadSTARD (Radiology Standards for Reporting of Diagnostic Accuracy Studies)

This systematic reporting tool was developed with radiological experts from every imaging domain of radiology. Although many items of the STARD tool have been deemed necessary when reviewing the quality of diagnostic accuracy trials, some items have been deemed not relevant and several items have been added to the list that are more specialized to the field of radiological sciences and in particular the reporting and interpreting of radiology diagnostic accuracy studies.

The name of this new tool is called the RadSTARD – Radiology Standards for Reporting of Diagnostic Accuracy Studies. This new tool was developed in Delphi fashion by studying the Standards for Diagnostic Accuracy (STARD) tool for radiologists to use to increase the quality of reporting diagnostic accuracy studies specific to radiology. The purpose of this elaboration document is to provide the radiologist and trainees with a supporting document for referral when utilizing the RadSTARD.

How to Use the RadSTARD

The following information for each item of the RadSTARD is provided below as an adjunctive to illustrate how to use the RadSTARD tool in that each item speaks to the value of the STARD but then branches off into further detail that is specific to radiology diagnostic accuracy studies. Although the STARD tool was developed to provide researchers guidance to increase the quality in reporting the results of their diagnostic accuracy studies; the purpose of the RadSTARD is to provide additional information for radiologists and their trainees. The RadSTARD is a guidance tool that was developed by radiologists to be used by radiologists and their trainees when developing their diagnostic accuracy studies specific to radiology. As well there are items within the RadSTARD that are in addition to the STARD tool whereas other items did not meet consensus by those who participating in the development of the RadSTARD.

Specifically the items that were not included were methods for calculating test reproducibility if done (STARD item # 13) and the provision of a cross tabulation of the results of the index test (including indeterminate and missing results) by the results of the reference standard; for continuous results, distribution of the test results by the results of the reference standard (STARD item # 19). Items where sentences are bolded are specific features to radiology diagnostic accuracy studies.

The RadSTARD checklist consists of 25 items. Additional items that met consensus by the working group included the recommendation for reporting the sample size and limitations, working with an imperfect reference standard and recommending reporting sound theoretical physics basis for the index test and reporting any modifications that occurred during the conduct of the radiology diagnostic accuracy trial.

The items in the RadSTARD are listed as per the following categories:

- title and abstract
- methods (patient eligibility and data collection)
- sample size
- diagnostic test methods
- imperfect reference standard
- test analysis

- results
- statistics
- discussion

A two page summary is provided as well which highlights the main features to include when either reporting the results of a radiology diagnostic accuracy study, or utilizing the RadSTARD when interpreting the results of published literature. Finally, it is encouraged that researchers use the RadSTARD when developing their diagnostic accuracy studies that are specific to radiology. To simplify the comparison to the STARD statement (tool) the STARD item number will be listed with each item of the RadSTARD in the two page summary depending where the item corresponds. Noting that two items of the STARD tool were not included as described above. Finally, the STARD is sometimes referred to as either the STARD statement or the STARD tool interchangeably. The RadSTARD reporting tool is only referred to as the RadSTARD or the reporting tool.

The information provided below describes and elaborates each item of the RadSTARD tool. In addition, for many items an example pertinent to radiology has been added for the user's reference.

TITLE & ABSTRACT **1.**	**Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract.**

The first item in the RadSTARD checklist recommends that radiology diagnostic accuracy studies are identified with the words "diagnostic accuracy" in the MeSH heading. The retrieval of diagnostic accuracy studies from electronic databases such as MEDLINE and EMBASE can be a daunting task due to the overwhelming amount of publications. As described by Wilczynski et al. (2013a) MEDLINE adds 10,000 to 20,000 references per week. Several years after the STARD was published in 2003 health science researchers from McMaster University worked to the develop specific search tragedies that would increase the success rates of retrieving diagnostic accuracy studies. Although they found no differences in MEDLINE between journals that endorsed STARD versus those that did not; there was some improvement noted in EMBASE (p = 0.02) (Wilczynski and Haynes, 2007a).

(Smidt et al., 2005) reported that to properly identify a diagnostic accuracy study, words such as 'sensitivity and specificity' or 'diagnostic accuracy' should be included to increase the success rate of the search. Although the use of sensitivity and specificity identified over 78% of the articles with this MeSH heading, only 100 of the total (686) articles were diagnostic accuracy studies. As studies were missed that were considered diagnostic accuracy studies, it is recommended by the cohort to use the term 'diagnostic accuracy' in the title and as a MeSH term when conducting a search.

Example: *"The diagnostic accuracy of magnetic resonance imaging in detecting cervical involvement in endometrial cancer".*
(Nagar et al., 2006 p. 431).

2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition.**

When stating the research question for diagnostic accuracy studies as per the STARD statement it advises to define the overall aim of the study as estimating diagnostic accuracy or comparing the accuracy between different subgroups of participants (Bossuyt et al., 2003a). Certainly; when articles were searched prior to the publication of STARD in 2003, details regarding the overall aim of the studies published in 2000 and in 4 journals identified as high impact journals referred to testing of the index test to the reference standard and the target condition. However; this information was dispersed throughout the article (Smidt et al., 2005). Since publication of the STARD evidence of explicitly stating the research question with the purpose of the study is relevant and readily available.

Example: "*The objectives of this study were to investigate the accuracy of magnetic resonance imaging (MRI) in predicting the depth of myometrial invasion in the preoperative assessment of women with endometrial cancer and to quantify the impact of MRI as an adjunct to predicting patients requiring full surgical staging"* (McComiskey et al., 2012 p. 1020).

When reviewing the aim of radiology diagnostic accuracy trials, the RadSTARD recommends comparing the index test to the reference standard to diagnose a specific condition. The internal validity of an intervention or study test should be assessed. Within the methods section of radiological literature the performance and effectiveness of a particular test is routinely analyzed for all therapeutic interventions. In particular, it is important to assess if the intervention used was tested in the appropriate patient population was considered and whether the results could be extrapolated in clinical practice. Secondly, the aim of the study should define that the index test was evaluated with an independent gold standard or reference standard (Budovec and Kahn, 2010). The purpose of clearly stating that the aim of diagnostic accuracy studies is to compare the results of the index test to the reference standard is to diagnose a specific condition (Leeflang et al., 2009). This aids the reader in determining the probability that a patient has a specific condition based on the association when the results of the tests are compared with the reference standard (Knottnerus and Muris, 2003).

Example: *A radiology diagnostic accuracy study was conducted to determine if the technical principles of STIR (short-time-from-inversion inversion recovery in patients who had undergone cardiovascular magnetic resonance imaging after the acute and chronic phases of myocardial infarction. This diagnostic accuracy trial was performed to compare the accuracy of STIR to those who had undergone cardiovascular magnetic resonance (CMR) imaging in diagnosing myocardial edema in a group of non-selected patients who had suffered an acute myocardial infarction* (h-Ici et al., 2012).

3. METHODS **PATIENT ELIGIBILITY, DATA COLLECTION**	**Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. whether primary or secondary care) should be provided.**

Clearly describing the eligibility criterion of participants enrolled in a diagnostic accuracy study is relevant. Failure to adequately list the criterion chosen to enroll patients in the study prevents disseminating useful information such as which patients were excluded from the trial (Bossuyt et al., 2003a). In other words, the study population should represent patients routinely seen in clinical practice and therefore the test chosen as reporting of diagnostic accuracy trial would be relevant to clinicians treating similar patients (Riegelman, 2013).

When describing the study participants the STARD criteria recommends that along with a description of the eligibility criterion sufficient detail with respect to patient recruitment should also be included so that the reader can determine if the participants are truly representative of the study population that could benefit from the index test (Riegelman, 2013).

The RadSTARD makes similar recommendations with the caveat that we know if the participants studied were receiving primary or secondary care. As described by Black (1990), clearly understanding whom the study results applied to is the most relevant question when assessing the methodology of accuracy in a radiology study. In other words details about participants with the disease entity versus those without disease needs to be described as this will enable future extrapolation of the study findings to clinical practice (Black, 1990).

The clinical manifestation or anatomical extent of a particular disease defined the spectrum of the disease impacts diagnostic accuracy. For instance, magnetic resonance imaging may not be able to differentiate those with or without early (stage IA) cervical carcinoma in asymptomatic patients. Alternatively, magnetic resonance imaging may be able to diagnose more advanced stages (IIB to IVB) of cervical cancer compared to those with a normal cervix. Therefore; the diagnostic accuracy of a test must be evaluated according to the evolution of a disease rather than a static entity (Black, 1990).

4. Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.

With respect to trial recruitment the STARD statement recommends providing information on recruitment of patients that is based on presenting symptoms, previous investigations or results obtained from either the index test or reference standard (Bossuyt et al., 2003b). Ultimately, the provision of considerable detail is required to determine if the patients included in the trial not only met eligibility criteria but it is also relevant to know if they were appropriately chosen for the index test under study (Riegelman, 2013).

Similar to the STARD the RadSTARD recommends trial recruitment should include both presenting symptoms and results from previous tests however; these items ranked lower priority than whether the participant received either the index test or reference standard. For example, a study comparing the diagnostic accuracy of metal artifact reduction sequence magnetic resonance imaging (MARS MRI) and ultrasound to detect painful metal on metal arthroplasty was recently published. For this particular diagnostic accuracy study participant recruitment included patients meeting the inclusion criterion as well as having undergone a MARS MRI within the previous year and who then subsequently went on to be protocolled to have an (ultrasound scanning) USS (index test). Whereas; newer patients who presented with pain were also included to receive both the prospective MARS MRI and USS (Siddiqui et al., 2014).

The assigning of participants into a diagnostic accuracy study can occur in one of the three following ways:

1. Participants who fulfill the eligibility criterion and target population of study are identified prior to undergoing the index test and reference standard.

2. Participants with and without disease are identified based on the results of their reference standard and who then undergo the index test.
3. Participants who have undergone the index test are identified to then undergo the reference standard.

In these three different scenarios, option one is considered the best method for patient recruitment as those identified represent the target population. Alternatively, option two and three may also be used however; each can result in biases from either of these methods. In option two if the investigator chooses only those who meet a clear diagnosis of the disease rather than including those in the gray area such as those with concomitant co-morbidities the results may result in spectrum bias due to the exclusion of those with co-morbidities and may have also have had a positive index test. Whereas; option three may also result in bias as the identification of study participants was based on findings from the index test which can result in verification bias (Riegelman, 2013).

In an effort to reduce the incidence of bias the RadSTARD recommends including participants based on those meeting eligibility criterion of whether or not the patient had undergone either the index test or reference standard. To clarify; participants in diagnostic accuracy studies undergo testing with both the index test and reference standard to compare the results. However; for this particular item of the RadSTARD it is recommending to include participants based on whether or not they had received the index test or reference standard pending the medical condition under study.

5. Was patient selection consecutive or non-consecutive?

The fifth item of STARD and the RadSTARD are similar when making recommendations of the sampling of study participants. For example, participants may be consecutively enrolled into the trial or perhaps they were enrolled after the index test or the reference standard were performed (Bossuyt et al., 2003b).

When reporting the value of diagnostic tests in diagnostic accuracy studies comparing the index test with the reference standard are typically reported in a consecutive cohort that is characteristic of the patient population under study (Van den Bruel et al., 2006). Although adopting to enroll consecutive study participants is a practical solution when testing a new method rendering a clinical diagnosis there are a number of concerns with this practice. For example choosing study participants consecutively from a local institution one may be influenced by the following factors:

- the geographical territory where the testing occurred
- the potential time interval during the year when the testing occurred (this is relevant when testing for seasonal conditions)
- selection of patient population (the ratio of in-patients versus outpatients within the sample)
- the selection process as per the referring physician

Therefore; when including participants in a diagnostic accuracy study the method by which the patients are included can render some degree of cohort bias. Therefore; it is very important that the methods section of the diagnostic accuracy article provide the following:

- a description of the demographic location where the study occurred with explicit details for the eligibility criterion
- for those within a consecutive cohort who were not included information detailing their exclusion should be provided as per the following illustrations:
- participant declined enrolment (provide reason)
- for those who did participate but whom did not undergo the index test (provide a reason)
- provide the list of cases that were not included due to the lack of proper examination (poor image quality due to artifact activity)

- the number of participants whose results were neither positive nor negative (indeterminate findings)
- the number of participants who did not undergo the reference standard (provide the reasons) the number of participants that were not assessed as the reference standard could not be classified (provide the reason) (Sardanelli and Di Leo, 2009a).

Ideally, patients are consecutively enrolled in diagnostic accuracy studies providing they have met the eligibility criterion within the specific study period and are considered appropriate to undergo the tests under consideration. This will render a spectrum of patients that would typically be seen within the clinicians practice. Not following this inclusion study pattern could impact the true prevalence of the target condition under study as well as the overall accuracy of the tests that are under study (Leeflang et al., 2009). In summary, although the RadSTARD recommends including participants in a consecutive series, it is also important to carefully review the characteristics of the patient population described.

6. Data collection (prospective or retrospective) should be provided with start dates and end dates.

Describing whether the study data was collected retrospectively or prospectively is an item that is recommended by the RadSTARD and is the sixth item of the STARD (Bossuyt et al., 2003b). However, for this particular item the RadSTARD also includes reporting the start and end dates of when the study was conducted. Although the start and end dates are also listed in the STARD tool it is combined in this particular item of the RadSTARD.

There are all types of bias that can affect the external validity of the study including whether or not the study was prospective versus retrospective. By way of illustration evaluating a newer technology against an older version of imaging technology will be biased. For example, comparing 64-row multi-detector computed tomography (MDCT) to 16-row MDCT to diagnose coronary stenosis may be deemed hemodynamically significant. For this example 16-row conventional coronary angiogram (CCA) was introduced in 2005 as the reference standard. Whereas; 64-row was introduced in 2007 as the reference standard for conventional coronary angiogram. The MDCT examinations were performed in both of these series of patients before the coronary angiogram was conducted. This rendered an overall impression that the image quality increased and that the number of non-diagnostic MDCT examinations was reduced. Therefore; it is not feasible to conduct a retrospective review comparing 16-row MDCT to 64-row MDCT in a retrospective series to detect significant coronary stenosis for the following reasons:

1. Lack of certainty that those who received the 16-row MDCT versus the 64-row MDCT were similar in nature with respect to the incidence of significant coronary artery disease. For example, if the participants had a high pretest probability of coronary artery disease this increases a potential bias with respect to individual sensitivity and positive predictive value to result in those who were tested with the 64-row MDCT.

2. If there were changes in the radiological team who worked with MDCT either before or after the start of the second series this could impact the diagnostic accuracy resulting in confound in favor of one series versus the second series of tests.

3. During the two years it took to learn how to use 64 row in MDCT would result in a bias in favor of 64 row in MDCT as it was conducted two years later.

4. As this was a four-year retrospective review the diagnostic accuracy for coronary conventional angiogram could have changed with respect to modifications and technology therefore; this would impact the reference standard between the two groups.

5. If those who perform the coronary conventional angiogram knew the results from the MDCT group incorporating the results could bias the interpretation of the reference standard conventional coronary angiogram (CCA). (Sardanelli and Di Leo, 2009a).

In summary, when interpreting the results of radiology diagnostic accuracy studies be it prospective or retrospective analyses it is important to consider when the technology was developed and carefully review how the results between the index test and reference standard were conducted and reported.

7. SAMPLE SIZE — Sample size and limitations should be provided.

This item is not included in the STARD statement.

When conducting a diagnostic accuracy trial, the results of the index test are compared to the results from the reference or gold standard test that each participant undergoes. Therefore, it is pertinent to know how many participants need to undergo both tests to render the results statistically significant. Knowing the sample size and its limitations will provide the confirmatory information that the results from the index test and the reference standard have statistical power (Riegelman, 2013).

Given that confidence intervals are routinely provided in the results of clinical trials and diagnostic accuracy studies this value can result in a larger sample size (Riegelman, 2013). If the sample size has a large positive population this will render increased precision around the confidence interval for sensitivity (Strassle et al., 2012). Determining whether the results of diagnostic accuracy studies have statistical power is very different from hypothesis testing investigations as there is no hypothesis that the index test is better than the reference standard. Determining what measurements around the confidence interval for sensitivity and specificity would be clinically relevant is appropriate. This is routinely done for hypothesis testing when tests are being compared to each other. In order to render a substantial significant power between the two tests this would require a large sample size. Sample sizes are calculated for case studies and may be used as a guide for diagnostic accuracies studies. For example, sample sizes of 100 to several hundred participants with or without the disease entity under study is an appropriate sample size for evaluating diagnostic tests. This is especially true when the pre-test probability of the disease entity among those who undergo the tests is greater than 50%. Alternatively, the sample size for a screening test would require a much larger sample size such as those found in cohort studies or randomized controlled studies (Riegelman, 2013).

If a diagnostic accuracy study was being done to determine if a newer imaging technique was better than a conventional (reference standard) test, it would be of interest to know if the difference between the two tests was statistically significant. If the difference between the two tests was not deemed to be significantly different this would clarify the fact that there was no clinically important difference rather than postulating if there may have been an important difference to note but it was not possible due to the lack of an appropriate sample size (Eng, 2004).

Sample size calculation should be performed when the protocol is being developed. In general, research studies that are properly designed include how statistical power was defined for the particular medical condition or disease entity under study. This includes calculating a sample size. Conversely, in most medical imaging journals the sample size calculation is not provided. The provision of the α error (0.05) is to be provided in a section on statistical analysis which implies the probability of false positive results. So if the results of the study render no statistical significance between the two groups not supplying a sample size and power analysis becomes an issue (Sardanelli and Di Leo, 2009a).

Appropriate Sample Size Parameters

An appropriate sample size is comprised of the following five study parameters:

1. the minimum expected difference which is also referred to as the effect size
2. the estimated measurement variability which is the standard deviation that is expected between measurements made amongst the interpreters
3. statistical power - as the desired study power increases the sample size also increases
4. the significance criterion which is typically set to 0.05
5. statistical analysis - one versus two tailed analysis (Eng, 2003).

Therefore, a sample size should be calculated apriori to be included with the study protocol whereby the definition for the sample size is derived by the minimal difference thought to render a clinical impact. This number is typically derived from a critical analysis of the literature by the radiologists in advance to conducting the study (Sardanelli and Di Leo, 2009a).

8. DIAGNOSTIC TEST METHODS	**Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale.**

The RadSTARD and the STARD both recommend reporting this item which is item 7 of the STARD. Providing the rationale for using a particular reference standard should also describe how it will answer the research question. Participants enrolled in the diagnostic accuracy study undergo both the index test and reference standard which is also considered the gold standard. The reference standard is the test or evaluation that is known and accepted amongst the medical community as the best method for confirming the diagnosis of a particular medical condition or disease entity (Riegelman, 2000).

Therefore; when reporting the results of radiology diagnostic accuracy studies, the performance of the index test is compared to the results from the reference standard or gold standard. For example, when conducting diagnostic accuracy studies in oncology the results of the diagnostic test under evaluation is typically compared to the pathological report which is the reference standard that is used to define the particular lesion under study. In this scenario radiologists and pathologists could be asked to provide dichotomous results (yes/no) when determining the presence of a disease lesion. The reference standard could be the pathological report which would state whether or not the disease was present. If the radiologists' interpretation matched that of the pathological report then this would be defined as a true positive. If the radiologists correctly diagnosed a non-malignant breast lesion as negative this would be a true negative. However; if the radiologist incorrectly diagnosed a non-malignant breast lesion as positive this would be a false positive. Finally, if the radiologist correctly diagnosed a malignant lesion as negative this would be a false negative. Data of this nature is typically presented in 2 x 2 contingency tables whereby the number of false positives, true negatives, false negatives and positives negatives are presented and can be found in the following table:

Two-by-two contingency table representing the results from the radiologists reports when compared to the pathologists reports

		Reference Standard	
		Positive	Negative
Radiological Exam	Positive	True positives (TP)	False Positives (FP)
	Negative	False negatives(FN)	True negatives(TN)

(Sardanelli and Di Leo, 2009a)

It is important to understand which statistical unit is being reported in radiology diagnostic accuracy studies as a participant with none or only one lesion renders no statistical significance. However if they have more than one such as those with liver metastases this can be a concern. In effect this can occur when studying the kidneys, breasts, lungs or any segment of the brain or coronary tree for example. Therefore; it is very important to report how the results were calculated, whether they were calculated by organ, segment, or lesion. The term ‘case’ should be avoided in scientific context. Instead report the statistical units that are under investigation for the diagnostic accuracy study (Sardanelli and Di Leo, 2009a).

9. IMPERFECT REFERENCE STANDARD	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified.**

This item is not provided in the STARD tool.

When conducting diagnostic accuracy studies the provision of a reference standard or gold standard test that is error-proof does not exist (Bertens et al., 2013). This sentiment was also expressed by the radiological experts who created the RadSTARD. Measures of accuracy frequently used include sensitivity specificity, likelihood ratios, diagnostic odds ratios or predictive values as they define how well the index test compares to the outcome of the reference standard (Reitsma et al., 2009). Therefore, alternatives to the reference standard can be provided with the provision that ample information about the imperfect reference standard is described.

Although multiple solutions have been proposed there is no universally accepted solution to replace a missing or imperfect reference standard when conducting diagnostic accuracy research (Reitsma et al., 2009). One alternative to an imperfect reference standard may be the use of a panel standard as the reference standard. Using a panel standard where multiples tests are evaluated is a viable alternative due to its approximate affinity to clinical practice. As there is no preferred method of utilizing panel standards it is important to discern how it was described (Bertens et al., 2013).

The following four key points should be considered.

1. **Constitution of Panel:** Reproducibility of the decision process made by the panel is more likely to occur if the same panel members are used. If this is not feasible then a researcher of particular expertise should be available to maintain consistency of opinion. If the panel consists of three or less interpreters then it is difficult to render a final disease classification. If the panel members are referred to as experts then a provision of their specific expertise and years of experience is suggested (Bertens et al., 2013).

2. **Thorough Description of Information Provided to the Panel: The domains of information presented to the panel could include items such as medical history, phys**ical exam and differential diagnoses. Whether or not the interpreter was blinded to this corresponding information needs to be stated (Bertens et al., 2013).

3. **The Panel’s Decision Process:** Classifying a disease as being present or not can be accomplished by the panel rating specific categories that render the certainty of diagnosis. Additional information can be recorded on the certainty of a final diagnosis which can enable subsequent analyses such as

a weighted analysis. Otherwise the initial decision process can be quite complicated by several decisions that must be made. Individual assessments are performed before meeting with the other panel members. Decisions are made according to pre-specified decision roles that have been agreed upon by the panel members. Classification of the target condition is usually presented as absent or present whereby the handling of disagreements is often described as a review by a plenary discussion or the provision of an additional expert for their opinion (Bertens et al., 2013).

4. **Was the Panel's Diagnosis Valid:** Reproducibility of the panel's decision is not typically performed. Inter-rater agreement amongst the panel members is accomplished by re-evaluating a sample of the study population whereby the panel is blinded to the initial diagnosis and compared for agreement. Validity of the panel diagnosis can be achieved by comparing the panel's diagnosis to clinical follow-up or an alternate reference standard (Bertens et al., 2013).

Composite Reference Standard or Latent Class Analyses

Utilizing a composite reference standard or latent class analyses is another alternative to an absent or imperfect reference standard. A composite reference standard is accomplished by combining multiple test results as per a predetermined algorithm which is used as a decision rule when diagnosing a particular disease as being present or not (Bertens et al., 2013). Composite reference standards have better discriminatory features than the components of reference standards in isolation (Reitsma et al., 2009). A statistical method called latent class analysis can also be used whereby the probability of a disease is deemed present based on the information available by the linking of multiple tests. These results are difficult to interpret as disease states are typically expressed as dichotomous (absent or not), rather than based on probabilities (Bertens et al., 2013, Reitsma et al., 2009).

Intrinsic Reference Standard

As intrinsic reference standard error occurs when a target condition does not produce the markers needed to confirm a diagnosis. In a radiology diagnostic accuracy trial, an example of an intrinsic reference standard error would be missing a tumor that was too small to be diagnosed with diagnostic imaging due to the level of perfusion employed (Reitsma et al., 2009).

Solutions for Imperfect Reference Standards

Main Classification	Main Features
Adjust for missing data for the reference standard	Attaining estimates of accuracy is done by observing patterns of agreement from other data and by comparing the results for both the index test and reference standard
Correcting an Imperfect Reference Standard	Correction for the estimates of accuracy for an imperfect reference standard are achieved by conducting a sensitivity analysis
Construct Reference Standard	Verification is done with the use of multiple test results to create a construct reference standard which is predetermined – referred to as composite reference standard. Alternatively a panel standard (expert opinion) or latent analysis (statistical measure) can be used instead.

Validation of the Index Test Results	The overall merit for a test is analyzing the results from the index test to other test findings

(Reitsma et al., 2009)

Irrespective of which alternate reference standard is chosen it is imperative that a description of the experimental procedures are adequately described rendering replication which is critical to understanding the journal results as published (Carp, 2012).

10. TEST Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies.

Providing details on the technical specifications for the index test and reference standard is the 8th item recommended by the STARD and also the RadSTARD. An adequate description when performing the index test is recommended so the reviewer can replicate the study as well as ascertain whether or not the index test can easily be facilitated within their clinical practices whereby discrepancies could inhibit similar results of diagnostic accuracy (Bossuyt et al., 2003b).

In addition, sufficient details should be provided on the technical aspects for the measures taken for the index test and reference standard. Plus the level of expertise for the evaluator should be detailed. With respect to the reference standard, information needs to be provided to justify its use for diagnosing the particular condition under study. The choice for the reference standard is not an easy decision to make for in order to compare the reference standard to the index test, the reference standard has to be able to confirm whether a disease is present or not (Riegelman, 2000).

In order to simplify this goal sometimes invasive tests such as biopsies are chosen as the reference standard (Riegelman, 2000). Although histopathology is the classic reference standard sometimes this choice of reference standard is not ethically possible. For asymptomatic subjects in screening programs negative examinations are compared to clinical and follow-up imaging exams. Whereas; in other situations the diagnosis is confirmed with an alternate imaging modality accepted as the standard of care for the medical disease entity under study. In studies where multiple lesions may be present in the same organ or segment of the body the topographic correlation between the histopathologic reference standard and imaging findings can be of concern. Therefore; the methods for correlation between radiology and pathology must be clearly described (Sardanelli and Di Leo, 2009a).

11. Sound theoretical physics basis of the index test should be provided for new techniques.

Recommending the provision of sound theoretical physics basis for the index test for new techniques is an item unique to the RadSTARD and not described by STARD. Understanding the theoretical physics basis of the index test is pertinent to all x-ray imaging modalities. The factors that affect the dose of radiation that patients are exposed to include factors such as the size of the patient, collimation (parallel light rays), filtration, beam energy and image processing. The highest peak kilovoltage is the most important reference in determining the acceptable patient dose for conventional radiography examinations. Digital radiography renders a broader range of digital imaging exposures than conventional radiography. However; it is important for the interpreter to be aware of the subtle differences used with digital systems to minimize the risk of increased radiation dose to the patient. For example, factors that affect the dose of radiation for mammography include small differences in the beam energy, screen film combinations, attention to breast tissue thickness and composition, varying standards for the grids, optical density and magnification. The tube current and peak kilovoltage are key elements to minimize the dose of fluoroscopy, more important features include collimation, source of the radiation dose to the skin, and distances of the patient to the image. Although there is less patient dose for computerized tomography versus conventional radiography with smaller primary doses delivered to the patient the dose calculations must be accounted for with exposure to adjacent tissue sections. With respect to fetal exposure and fetal dose effects numerous variables also need to be monitored such as knowing the size and depth of the fetus and the relation of the x-ray tube in reference to the orientation of the patient (Parry et al., 1999).

The following is an example of the detail that should be provided when reproducing the analysis of functional MRI (fMRI). In an effort to ensure that the reader comprehend the model being used it is pertinent to describe the approach in detail. Typically fMRI analysis are reported using the general linear model (GLM) whereby there are various differences between the models that should be defined. Most of these differences are apparent upon analysis with the software packages. Therefore; rationale for the software parameters utilized such as how the covariance structure was modelled should be thoroughly described. When comparing the results from individuals studied, provision of the precise statistical tests utilized to draw inferences should also be included (Poldrack et al., 2008).

12. Modifications during the study should be reported if they occurred.

Reporting any modifications that occurred during the conduct of a radiology diagnostic accuracy study is recommended by the RadSTARD. This item is not described with the STARD tool. As there are reportedly many potential sources of bias in radiologic studies; any modifications during the conduct of a radiology diagnostic accuracy study could bias any aspect of the study. For example, interpreting the result of the index test index must be done without knowledge of the results from the reference standard. Similarly; the reference standard results must be interpreted without knowing the results from the index test (Obuchowski, 2003b).

If the index test or the reference standard is interpreted without the proper blinding this is referred to as review bias. An example of this would be a diagnostic accuracy study comparing the accuracy of computerized tomography (CT) and ultrasound to diagnose tumors. When the ultrasound is being conducted, neither the radiologist nor the technician should be aware of the results from the computerized tomography as this may result in the technician over scrutinizing a certain area where a tumor was localized. Similarly; the radiologist could over-interpret the results of the lesion from the ultrasound if the results of the CT were known. The easiest way to prevent this bias from occurring is to blind the technician and the radiologist from the other test results (Obuchowski, 2003b).

13. Radiology Diagnostic Accuracy Trials should report cut-off values for specific diagnostic criteria for index and reference tests.

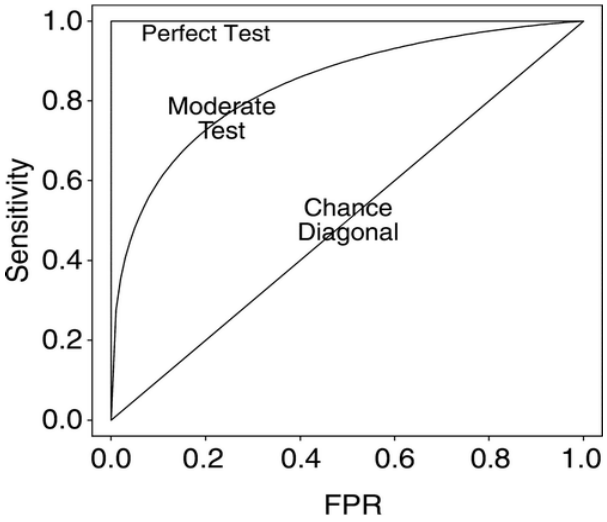
ANALYSIS

The STARD tool (item 9) recommends that the reference standard and index test are clearly described, including the rationale for their use and whether cut-offs in the results were provided as this differentiates positive from negative findings. It is important that the investigators define their categories of results a priori or when the study results were obtained. Replicating the study findings would be difficult to achieve if the authors chose their cut-off values after the tests were completed (Bossuyt et al., 2003a, Riegelman, 2000).

The ability of a test result to diagnose a condition as either positive or negative is not always straightforward as it is to determine if a test is positive or negative due to the presence of an organism or a drug administered. In situations where tests provide numerical data the investigators need to define the cut-off values that differentiate positive from negative findings (Riegelman, 2000).

Similarly, the RadSTARD recommends reporting cut-off values but in addition it recommends reporting these cut-off values or cut points for specific diagnostic criterion for both the index test and reference standard. The receiver operating characteristic curve (ROC) is one of the most popular methods of measuring accuracy of a diagnostic test that is measured by the area under the curve. The value for the ROC curve can range from 0.0 and 1.0 where a ROC curve of 1.0 is considered perfect accuracy and test 0.0 is considered perfectly inaccurate (Obuchowski, 2003a).

The following is an example of the application for the ROC curve in radiology research. Patients with the suspected diagnosis of multiple sclerosis (MS) were examined with head MRI and CT. The images were scored independently by the radiologists who were unaware of the patient's final diagnosis. They were scored as definite MS, probable MS, possible MS, probably MS and definitely not showing MS. The reference standard consisted of an expert panel specializing in MS, where the results of follow-up tests six months later with other diagnostic tests were compared. However; the results from MRI and CT were not included to minimize bias (Obuchowski, 2003a).



“Graph shows comparison of three ROC curves. A perfect test has an area under the ROC curve of 1.0. The chance diagonal has an ROC area of 0.5. Tests with some discriminating ability have ROC areas between these two extremes” (Obuchowski, 2003a p.6).

The results of MR imaging revealed a ROC curve of 0.82 which is good but not a definite diagnosis. Alternatively, the results of CT imaging revealed a ROC curve of 0.52 meaning that the results from CT were no more accurate than guessing the diagnosis of MS. The researchers in this study concluded that MR imaging is required to diagnose MS however; they also advised that normal MR imaging does not necessarily exclude a diagnosis of MS (Obuchowski, 2003a).

14.	**The training and number of investigators should be described including any extra training for new techniques.**

With respect to interpretation of the index test and reference standard the training and number of investigators should be described as recommended by the STARD and the RadSTARD. As reader variability has been found within the field of diagnostic imaging, the level of training could impact whether similar results can be found in clinical practice of varying degrees of experience (Bossuyt et al., 2003b, Riegelman, 2013). In addition, the RadSTARD recommends including a description of any extra training for new techniques.

The provision of additional factors added to a checklist for distinct classes of diagnostic imaging tests has been recommended. For example, with assessing the accuracy of medical imaging modalities, certain items on the checklists may differ according to their technological stage of development when the study was performed. Likewise, the clinical context or variability amongst interpreters may be more significant pending the study stage when aspects of technical parameters are being established for new imaging modalities (Gatsonis, 2003).

15. **Whether readers were blinded to prior test results or clinical information should be known.**

Whether the interpreters of the index test or the reference standard were blinded to the results of either prior study results or clinical information should be described. Essentially, if those interpreting results of the index test were not blinded to the results obtained from the reference standard the interpretation of results could be inflated affecting the overall extent of diagnostic accuracy (Bossuyt et al., 2003a). Both the STARD (item 11) and the RadSTARD are in agreement with this recommendation.

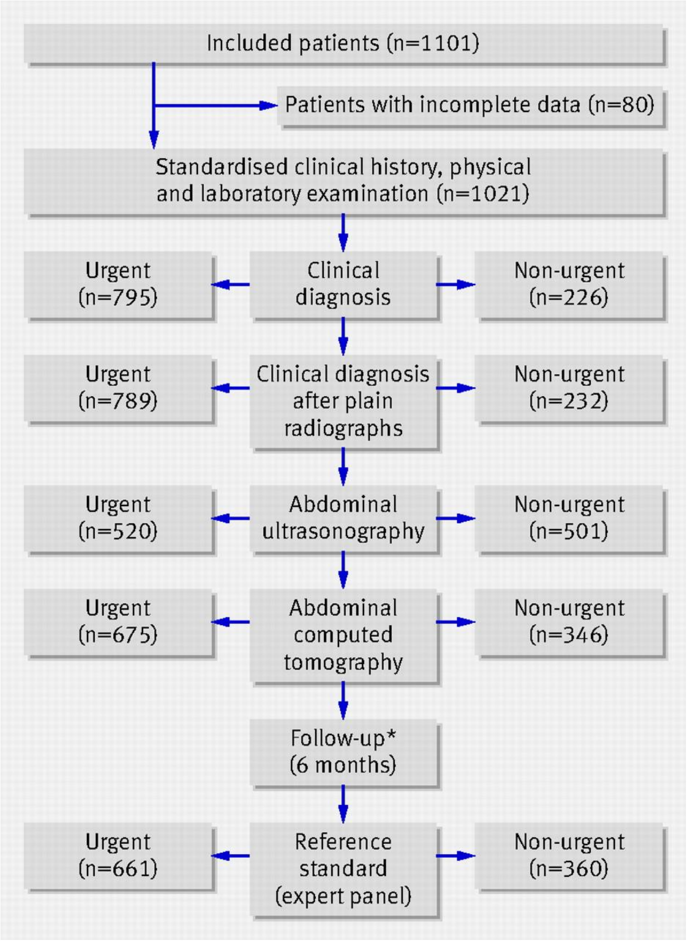
Determining how patients are selected for evaluation influences how well the results reflect the true accuracy of the test for the study population. Whether accuracy is reported as sensitivity and specificity with a numerical value or implicitly through diagnostic imaging interpretation almost all processes of diagnostic evaluation are biased to some degree. To avoid test review bias the radiologist should be blinded to other contributory findings as this will ensure that the diagnosis was made entirely on the study results. Such interpretation will minimize the possibility of overestimating the degree of sensitivity and specificity (Black, 1990).

As an example when a radiologist attempts to differentiate hemangioma of the liver versus metastatic disease via magnetic resonance imaging, certain diagnostic criterion must be met. If prior knowledge of the diagnosis was known, subtle ambiguities such as the contour and shape of the liver as well as relative signal intensity on magnetic resonance imaging could impact the radiologists interpretation of the lesion (Black, 1990). In other words, if the interpreter knows the results for the index test and other clinical results when interpreting the reference standard for the study cohort this could inflate the degree of diagnostic accuracy. By thoroughly and accurately reporting diagnostic accuracy studies the incidence of bias can be minimized and generalizability of results can be assessed (Smidt et al., 2005).

16. RESULTS **Radiology Diagnostic Accuracy Trials should report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations.**

STARD item 16 recommends the insertion of a flow diagram to illustrate the number of eligible participants to participate in the trial, the interventions that were administered, study outcomes and final results area (Bossuyt et al., 2003a). The use of a flow diagram was first initiated with the CONSORT (Consolidated Standards of Reporting Trials) statement and it resulted in improving the quality of reporting for randomized controlled trials (Smidt et al., 2005). By including a flowchart it also illustrates which patients were chosen to participate in the trial including those that were excluded plus it helps investigators understand the participant characteristics of the study population (Riegelman, 2013).

The provision of a flow diagram was also recommended by the RadSTARD. A diagnostic accuracy study was conducted on patients presenting to the emergency department with acute abdominal pain which evaluated the advantage of adding plain radiographs, abdominal ultrasound and computed tomography after the patient had been examined. The accuracy of these exams was compared in a non-selected group of consecutive patients that presented with non-traumatic abdominal pain. The following is an example of the flow diagram that was provided for this study (Laméris et al., 2009).


Included patients (n=1101)
Patients with incomplete data (n=80)
Standardised clinical history, physical and laboratory examination (n=1021)
Urgent (n=795)
Clinical diagnosis
Non-urgent (n=226)
Urgent (n=789)
Clinical diagnosis after plain radiographs
Non-urgent (n=232)
Urgent (n=520)
Abdominal ultrasonography
Non-urgent (n=501)
Urgent (n=675)
Abdominal computed tomography
Non-urgent (n=346)
Follow-up* (6 months)
Urgent (n=661)
Reference standard (expert panel)
Non-urgent (n=360)

(Laméris et al., 2009 p. 2)

17. **Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies should be provided.**

Providing details on the patient demographics (item 15 of the STARD) is also recommended by the RadSTARD. When the study population is adequately described it enables the reviewer to decide if the trial would be applicable for their clinical practice setting (Bossuyt et al., 2003b). For knowledge translation to occur to clinical practice based on the test results from a research study the population studied needs to be representative of the population in whom the results will be applied. Regardless, selection bias is difficult to avoid as it may occur due to the overall context of the trial or study design. Nonetheless, authors must acknowledge the incidence of selection bias within the discussion section of their article (Sardanelli and Di Leo, 2009a).

Diagnostic modalities are interventions that have non-negligible risks such as exposure to ionizing radiation and injection of contrast material which render a greater propensity of the test being performed in study participants that are suspected of already having been diagnosed with the disease under study. This is referred to as diagnostic safety bias. The provision of any co-treatments or therapies for any of their diagnostic imaging can inhibit enrollment or reflect changes in the radiological findings for the study sample. In particular, the internal validity of the study will be impacted if only some of the study sample receive additional co-therapies (Sardanelli and Di Leo, 2009a).

This is different in spectrum bias which occurs when the prevalence of disease type (acute or chronic) or stage of tumor growth has occurred within the study sample as it is different from those who will undergo the tests or treatment later in clinical practice. Population bias occurs when there is an unusually increased propensity for the disease entity to be prevalent in the study sample. However; there are times that spectrum bias is applicable such as when the performance or diagnostic accuracy of a new imaging modality is being evaluated in a given disease state. Even if the results are overwhelmingly positive, the use of the tests can't be transferred to clinical practice until further testing is done. Finally, the presence of co-morbidities can affect the study sample as the results of diagnostic imaging will be impacted (Sardanelli and Di Leo, 2009a).

In summary, in an effort to minimize the incidence any one of these biases from occurring and over estimating the diagnostic accuracy results, it is paramount that the demographics of the patient population be thoroughly described.

18. The severity (spectrum) of the disease entity should be explicitly reported.

STARD item 18 advocates reporting the severity of the target condition within the study population as the presence of co-morbid conditions or duration of the disease can affect the overall diagnostic accuracy measures (Bossuyt et al., 2003a). This item is also recommended by the RadSTARD.

The severity or level of a disease that a patient has will also impact the probability of a test result pending the degree of severity of the disease. In general the severity of the disease should be lower in patients who are randomly diagnosed as having a disease versus those in whom the disease has been present for some time. Therefore; the level of disease severity would be lower in patients who were diagnosed upon routine screening versus patients who were previously diagnosed by their clinical practitioner. As such, this will render a direct impact on the level of sensitivity and specificity which would be expected to be higher in symptomatic patients versus asymptomatic patients who were diagnosed with earlier stages of the disease. This level of difference can be seen in screening programs for oncology whereby tumors are being detected at earlier stages or later stages of treatment. Essentially those chosen to be included in the study is reflected by the overall severity of disease as this influences the degree of sensitivity and specificity from the study findings (Sardanelli and Di Leo, 2009a).

19. The time difference between the index test and reference test and details of any other treatments provided between the two tests should be provided.

The time interval that it took to perform the index test and the reference standard is item number 17 of the STARD recommendation tool. The reason for including this information is if there was a time lapse in the patient's condition this could change and affect the target condition being studied (Bossuyt et al., 2003a). In addition, the STARD tool recommends reporting any treatment or investigations that may have occurred between the performance of the index test and reference standard or estimates that impacted the overall results. It is recommended that the index test and reference standard be performed within a short interval of time as the administration of any intervention or treatment between the two tests may affect the ability of the second test to diagnose the disease under study. Therefore, when conducting and/or interpreting diagnostic accuracy studies providing information on the assignment of how patients were recruited with respect to the time interval for receiving or conducting the index test and reference standard should be included (Riegelman, 2000).

The RadSTARD also recommends reporting the time interval between the index test and reference standard and the provision of any additional treatment for the same reasons as recommended by the STARD. However, in addition when radiologists are interpreting results of the two tests reading-order bias can occur. Suppose a patient undergoes two tests - treatment A and treatment B. If they are read by the same interpreter the images read last (treatment B) by the radiologists will be interpreted more accurately than the images that were retained from treatment A (Obuchowski, 2003b).

Typically when conducting diagnostic accuracy studies one is interpreting the results of the index test after the performance of the reference standard whereby the method of interpretation is done by two separate readers. However, this type of reading-order bias is mentioned as this can occur in other types of radiology research (Obuchowski, 2003b). It is therefore recommended by the RadSTARD for the reader to be cognizant that this type of bias did not occur.

Assessing the diagnostic accuracy for the index test involves verifying the results from the index test to that of the reference standard for each participant studied. If not all participants are verified or some participants are verified by additional reference standards this can result in verification bias. Whereas; partial verification bias occurs when not all study participants undergo the reference standard resulting in an over-estimation of diagnostic accuracy (Leeflang et al., 2009).

20.	**Adverse events should be reported for either the index test or reference standard.**

STARD item 20 recommends reporting any adverse events that occurred when the index test or reference standard were performed. This item is of particular interest if the reference standard was invasive as the goal is to minimize any element of risk in future studies. Therefore; all adverse events should be reported (Bossuyt et al., 2003a). Safety data including the incidence of any side effects or adverse events occurring needs to be reported in sufficient detail thereby illustrating the scenario and time interval when the adverse event occurred (Riegelman, 2000).

The RadSTARD also recommends reporting any adverse events that may have occurred from the index test or reference standard. In radiology research, an additional consideration is to avoid the exposure to any unnecessary ionizing radiation or as “low as reasonably achievable” which is commonly referred to as the ALARA principle (Sardanelli et al., 2010 p. 6).

21. **STATISTICS**	**Radiology Diagnostic Accuracy Trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI).**

This item determines whether estimates of diagnostic accuracy were reported as recommended by the STARD checklist. Receiver operator characteristic curves with 95% confidence intervals are recommended to illustrate the sensitivity and specificity of the index test to the reference standard (Bossuyt et al., 2003a, Riegelman, 2000).

The RadSTARD also recommends stating measures of diagnostic accuracy and uncertainty with p-values and confidence intervals of 95%. In order to assess the accuracy of study results from a diagnostic accuracy study the confidence interval must be calculated along with an estimate that can be found between the results of two tests with a specified degree of certainty (95%, 99%) (Villar, 2011).

Therefore, the confidence interval is comprised of a range of values that are believed to represent the true population value. Although this population value is not known it can be estimated from a sample size that has been appropriately chosen. Calculated confidence intervals define how precise the estimates are for the population chosen under study (Medina and Zurakowski, 2003).

However, if the sample size is limited this will result in a sample estimation that is not a true representation of study populations and will therefore be expressed with a wide confidence interval. The wider the confidence intervals are the less precise the estimation is rendering doubt in the reliability from the observed results. By way of example, if the specificity of a particular diagnostic modality to diagnose a given disease entity resulted in a value of .75 and a confidence interval equal to [0.57, 0.93], given the wide confidence interval there is a greater chance to underestimate or overestimate the specificity of the test. Confidence interval measurements are impacted by the sample size and if there was any variability in the sample. They do not define whether or not there were any possible errors in the design of the study with respect to implementation and statistical analysis (Sardanelli and Di Leo, 2009a).

22. Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers for the index test and reference standard; and describe how this data was handled.

Reporting how the data was handled for indeterminate or missing results and outliers is an item from the RadSTARD that is similar to the STARD however; the STARD specifically recommends reporting indeterminate results from the index test as ignoring such results can bias measures of diagnostic accuracy (Riegelman, 2000, Bossuyt et al., 2003b).

The RadSTARD recommends reporting indeterminate/missing results and outliers for all data which includes the index test and the reference standard. Upon successfully meeting the eligibility criteria, participants are expected to undergo testing of the index test and reference standard. For those who undergo the index test and not the reference standard, this loss of follow-up can lead to bias when compared to those who remained in the study (Riegelman, 2000).

There are instances when test results are uninterruptable such as an insufficient specimen from a needle biopsy, abdominal gas affecting the interpretation of a pelvic ultrasound scan, or dense breast tissue when conducting mammography screening. It is critical that all cases from this study are reported when analyzing the results of the study. This includes reporting the frequency of uninterruptable results when comparing tests (Obuchowski, 2003b).

23. Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability including > 3 observers with variable expertise.

This item of the RadSTARD is similar to STARD (item 23) which suggests that estimates of variability of diagnostic accuracy should be reported between subgroups if done. If a subgroup analysis was planned it should be conducted apriori to the initiation of the study as this will determine if there is heterogeneity present in the final analysis (Bossuyt et al., 2003b).

Although previous authors have analyzed the quality of reporting of diagnostic accuracy literature with a modified STARD tool that did not include this item, the RadSTARD does recommend including it (Walther et al., 2014, Wilczynski, 2008). Alternatively, Gatsonis (2003) purports that most items of the STARD do pertain to diagnostic imaging including the requirement to report the degree of inter-observer variability between readers. Assessment of this feature is very pertinent to diagnostic imaging which also encourages that any subgroup analysis also be reported by the cohort.

Whenever an observer reports the results of the test, there is a potential for either interobserver or intraobserver variability to occur. For example if two radiologists provided different interpretations for the same digital image it is called interobserver variation. Whereas; if an interpreter reported any significant difference between two different readings throughout the day this would be referred to as intraobserver variation (Riegelman, 2000). The RadSTARD recommends including > 3 observers of variable experience when interpreting the study results.

Tests that produced the same results when replicated imply precision. In studies that examine reproducibility study results are interpreted at least twice. Tests that are reproducible render the same results when interpreted by either the same examiner twice or by others who are unaware of each other's findings. This is commonly referred to as interobserver and intraobserver reproducibility. Evaluation of interobserver reproducibility occurs when the second interpreter records the results without knowing the results obtained from the first interpreter. When the same interpreter records their results twice this results in evaluating intraobserver error as the interpreter is unaware of his or her results from the initial reading (Riegelman, 2013).

Example: A systematic review conducted by Roposch et al. (2006) analyzed the quality of diagnostic accuracy studies that utilized ultrasonography (US) in diagnosing developmental dysplasia (DDH) of the hip. The purpose of the review was to determine the overall quality of reporting for diagnostic accuracy studies that used US studies to diagnose DDH. Studies that were eligible included those that measured the diagnostic accuracy of ultrasound methods reported in detecting DDH in neonates, infants and older children. In particular, studies were included that described how the ultrasound exam was conducted and interpreted. In addition, studies that involved reproducibility rendering reliability and validity were also evaluated (Roposch et al., 2006).

Of note, the quality of reporting for this STARD item was in general poorly done. Only seven studies were identified for review whereby the authors described validity of different US methods and only four studies assessed reliability. With respect to validity of US for the hip the appropriate image acquisition was provided. However, estimates of reproducibility were reported as continuous variables with means and standard deviations and no accounting for agreement between the two raters. Unless the methods are sound, interpretation and generalizability from these results cannot be concluded (Roposch et al., 2006).

The following diagram defines how to ascertain if a test is useful:

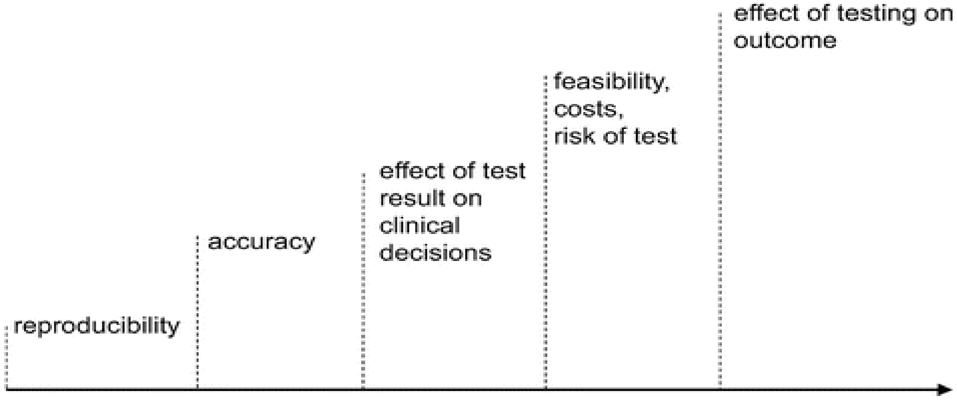

effect of testing on outcome
feasibility, costs, risk of test
effect of test result on clinical decisions
accuracy
reproducibility

Figure 1 – “The first step in determining if a test is useful is reproducibility or reliability. Reproducibility of a test is absolutely necessary. The following step is accuracy of the test which is defined by its validity and reliability. When these measurement properties are provided diagnostic yield of the test is useful in making further clinical and treatment management decisions. For the purpose of this review and example, the first two steps combined render accuracy” (Roposch et al., 2006 p. 856).

24. Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.

Both the STARD and the RadSTARD recommend reporting the estimates of test reproducibility. This is commonly conducted with the coefficient of variation and Kappa statistics (Bossuyt et al., 2003b). Reproducibility of study results may also be described as reliability meaning that the same results would be achieved if the test were repeated again. Although the STARD criterion recommend reporting quantitative assessments for reproducibility it does not require reporting the conduct of the methods utilized for how the estimates of reproducibility were performed (Riegelman, 2000).

This item corresponds to item 24 of the STARD. Reporting estimates of reproducibility is commonly conducted with the coefficient of variation and kappa statistics (Bossuyt et al., 2003b).
Kappa or Cohen k statistic is interpreted as follows:

1. A k value between 0.41-0.60 represents moderate agreement amongst reviewers
2. A k value between 0.61-0.80 represents substantial agreement and
3. A k value between 0.81-1.00 renders almost perfect agreement amongst reviewers

(Walther et al., 2014)

Providing estimates of test reproducibility is another item of the current STARD tool that previous authors did not include when they conducted reviews of the use of the STARD tool when reporting diagnostic accuracy study results (Walther et al., 2014, Wilczynski, 2008). However, the RadSTARD does recommend including this item when reporting the results of radiology diagnostic accuracy studies. When describing the results for the diagnostic performance of any imaging modality, it is being compared to the reference standard. So the question is if the test was repeated n times what would be the probability that the same result would be achieved (Sardanelli and Di Leo, 2009b)?

In most scenarios a radiologist records the value of a variable only once as opposed to repeating the measurement to enhance estimate precision. An example could be the measurement of left ventricular function using cardiac magnetic imaging resonance imaging (MRI) cine sequence in a patient post-ischemic event. By measuring the volumes for both the diastolic and systolic phases the ejection fraction can be calculated (Pope, 2009).

To conduct this type of analysis a software package would render the radiologist the ability to delineate the surface of endocardial tissue in all slices to measure the heart mass based on the amount of ventricular blood in the systolic and diastolic phases. The software program for this application would render suboptimal results as it would utilize algorithms to try to get around each curve for the relative cardiac contour. Manual correction would be necessitated by the radiologists to correct for the interpretation from the software package. As this second step requires choosing slices in cardiac phases for segmentation the radiologist or observer introduces variability in the measurements. As it would be nearly impossible to obtain the exact same measurement repetition by the same observer would result in different values (Pope, 2009).

This is known as intra-observer variability whereby variability in responses occurs when an observer repeats the same measure in the same clinical scenario more than two or three times. Conversely, when the measurement is performed by two or more observers the variability in results is even more pronounced and is known as interobserver variability. Although reproducibility and variability are used interchangeably for the purposes of the statistical explanations variability will be used (Pope, 2009).

The following example illustrates why measuring for interobserver and intraobserver variability is so important. Cengiz et al. (2014) conducted a retrospective review of over 200 abdominal ultrasounds that were conducted between 2010 to 2011 in patients with non-alcoholic liver disease (NAFLD). The purpose of this review was to measure the incidence of intraobserver and interobserver variability that resulted from sonographic examinations performed to diagnose the steastosis grades of NAFLD. In addition, liver enzymes (alanine aminotransferase (ALT) and aspartate aminotransferase (AST) plus enzyme ratio (AST/ALT) were also evaluated in reference to the degree of the hepatosteatosis. The hepatic images on ultrasound were graded as, mild, moderate or severe hepatic steatosis. These evaluations were repeated within one month's time.

Table 1: Intraobserver agreement rates

	Intraobserver agreement (kappa)	**Intraobserver agreement (percentage)**
Observer 1	κ=0.356	51%
Observer 2	κ=0.591	68%

(Cengiz et al., 2014 p. 5456).

Table 2: Interobserver agreement rates

	Interobserver agreement (kappa)	**Interobserver agreement (percentage)**
1st evaluation	*k*=0.208	39%
2nd evaluation	*k*=0.225	40%

(Cengiz et al., 2014 p. 5456).

Table 3: Number of individuals with elevated ALT and/or AST and with an elevated (> 1) AST/ALT ratio in each group

	Individuals with elevated ALT and/or AST (percentage)	**Individuals with elevated (> 1) AST/ALT ratio (percentage)**
Normal	3/27 (11%)	13/27 (48%)
Mild hepatic steatosis	8/29 (27%)	10/29 (34%)
Moderate hepatic steatosis	5/37 (13%)	11/37 (30%)
Severe hepatic steatosis	2/29 (10%)	6/20 (30%)

(Cengiz et al., 2014 p. 5457).

Statistical Measures:

For this particular study analysis of the intraobserver and interobserver variability were measured with Kappa statistics. The percentage of participants with high liver enzymes (ALT and/or AST or AST/ALT ratio over one was calculated for each group. The interobserver agreement for the first observer was 51% rendering a fair Kappa (K = 0.356) whereas the interobserver agreement for the second observer was 68% (K = 0.591). The results for the interobserver agreement between the first and second readings were 39% and 40% fair kappa of (K = 0.208) and (K = 0.225) respectively. Elevated liver enzymes were similar between the patient groups and correspondent the degree of the hepatosteatosis. The authors concluded that the visual evaluation for NFALD via ultrasound examination renders substantial intra-variability results which limits reproducibility of the findings (Cengiz et al., 2014).

When determining the degree of interobserver and interobserver variability between measurements; other alternatives to Kappa can be used. Although the Bland-Altman analysis which was originally created by two physicians to enable comparison in medicine of two methods of measurement it has since been extended by others for measuring the degree of intraobserver and interobserver variability for continuous variables (Pope, 2009).

A Bland-Altman analysis consists of a value that is expressed with the same units of measurement as the variable that is being measured resulting in direct interpretation. In terms of agreement between two observers (interobserver variability) the higher the agreement the lower the variability. Likewise, the greater the agreement a single observer has will result in a lower intraobserver variability (Pope, 2009).

When two measurements are being evaluated a certain lack of agreement is possible as both are indirectly measuring the same variable there is also the propensity for error. As clinicians and researchers are familiar using the reference standard in certain patient populations when evaluating it against a new measurement the amount of difference between the old and new measurements needs to be accurately compared. Bland Altman plots can be created to illustrate the level of agreement and/or disagreement between the two measures (Connelly, 2008).

25. DISCUSSION	The clinical relevance of the study findings should be provided.

Encouraging the discussion of clinical findings from the results of diagnostic accuracy studies is recommended by both the RadSTARD and STARD tool. Study limitations and or shortcomings that would be expected if the test was conducted in future studies should also be provided (Bossuyt et al., 2003a). In general, the discussion of the study results involves both the interpretation and comments based on the study findings (Riegelman, 2000).

Discussion of the study findings can be provided by two different ways. One approach would be to summarize the overall study findings. For example, in patients with colorectal cancer discussing the results of the study to diagnose liver metastases would involve demonstrating that contrast enhanced computerized tomography resulted in a higher degree of accuracy than ultrasound. Alternatively, the authors could approach their discussion by reflecting on arguments that may have been presented in the introduction of the study. In this particular example they could include discussing how the incidence of colorectal cancer is escalating and then present the results (Riegelman, 2000).

Overall; the point of the discussion section is for the investigators to comment on the results of their study which may include comparing and contrasting results from this study to the results of previous trials. Therefore; such differences between the trials would be accounted for both quantitatively and qualitatively. Finally, study limitations would also be included as this would illustrate how the investigators attempted to avoid any level of bias while conducting their study (Riegelman, 2000).

References

ABRAHAM, M. & PURKAYASTHA, B. 2012. Making a difference: Linking research and action in practice, pedagogy, and policy for social justice: Introduction. *Current Sociology,* 60**,** 123-141.

AFIFY, M. F. 2008. Action research: Solving real-world problems. *Tourism and Hospitality Research,* 8**,** 153-159.

AL-SULTTAN, F. M., FRAGKOS, K. C., BOGDANOS, D. P. & FORBES, A. 2012. Tu1224 A Systematic Review and Meta-Analysis of Pancreatic Autoantibody's (Pab) Diagnostic Accuracy vs Standard Diagnosis in Patients With Inflammatory Bowel Disease. *Gastroenterology,* 142**,** S-778-S-779.

ALSHAMARI, M., NORRMAN, E., GEIJER, M., JANSSON, K. & GEIJER, H. 2015. Diagnostic accuracy of low-dose CT compared with abdominal radiography in non-traumatic acute abdominal pain: prospective study and systematic review. *European radiology***,** 1-9.

ALTMAN, D. G. & MOHER, D. 2014. Importance of Transparent Reporting of Health Research. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

APPLEGATE, K. E. & CREWSON, P. E. 2002. An introduction to biostatistics. *Radiology,* 225**,** 318-322.

ASTIN, M. P., BRAZZELLI, M. G., FRASER, C. M., COUNSELL, C. E., NEEDHAM, G. & GRIMSHAW, J. M. 2008. Developing a sensitive search strategy in MEDLINE to retrieve studies on assessment of the diagnostic performance of imaging techniques. *Radiology,* 247**,** 365-373.

BACHMANN, L. M., TER RIET, G., WEBER, W. E. & KESSELS, A. G. 2009. Multivariable adjustments counteract spectrum and test review bias in accuracy studies. *Journal of clinical epidemiology,* 62**,** 357-361. e2.

BAILEY, B. & AMRE, D. K. 2005. A toxicologist's guide to studying diagnostic tests. *Clinical Toxicology,* 43**,** 171-179.

BALSIGER, P. W. 2004. Supradisciplinary research practices: history, objectives and rationale. *Futures,* 36**,** 407-421.

BANSAL, G. J. & YOUNG, P. 2015. Digital breast tomosynthesis within a symptomatic "one-stop breast clinic" for characterization of subtle findings. *The British journal of radiology,* 88**,** 20140855.

BARBOUR, R. S. 2001. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *Bmj,* 322**,** 1115-1117.

BARDOU, D., BOURSIER, J., CARTIER, V., LEBIGOT, J., MICHALAK, S., OBERTI, F., FOUCHARD-HUBERT, I., ROUSSELET, M. C., AUBE, C. & CALÈS, P. 2013. FIRST INTENTION-TO-DIAGNOSE COMPARISON OF ARFI AND FIBROSCAN IN CHRONIC LIVER DISEASES. *Journal of Hepatology,* 58**,** S7.

BERTENS, L. C. M., BROEKHUIZEN, B. D. L., NAAKTGEBOREN, C. A., RUTTEN, F. H., HOES, A. W., VAN MOURIK, Y., MOONS, K. G. M. & REITSMA, J. B. 2013. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS medicine,* 10**,** e1001531.

BLACK, W. C. 1990. How to evaluate the radiology literature. *American Journal of Roentgenology,* 154**,** 17-22.

BOHTE, A. E., DE NIET, A., JANSEN, L., BIPAT, S., NEDERVEEN, A. J., VERHEIJ, J., TERPSTRA, V., SINKUS, R., VAN NIEUWKERK, K. M. & DE KNEGT, R. J. 2014. Non-invasive evaluation of liver fibrosis: a comparison of ultrasound-based transient elastography and MR elastography in patients with viral hepatitis B and C. *European radiology,* 24**,** 638-648.

BOSSUYT, P. M. 2008a. STARD Statement: Still Room for Improvement in the Reporting of Diagnostic Accuracy Studies 1. *Radiology,* 248**,** 713-714.

BOSSUYT, P. M. 2009a. Diagnostic accuracy reporting guidelines should prescribe reporting, not modeling. *Journal of clinical epidemiology,* 62**,** 355-356.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L., LIJMER, J. G., MOHER, D., RENNIE, D. & DE VET, H. C. 2015a. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology,* 277**,** 826-832.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L., LIJMER, J. G., MOHER, D., RENNIE, D. & DE VET, H. C. 2015b. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *Radiology***,** 151516.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., LIJMER, J. G., MOHER, D., RENNIE, D., DE VET, H. C. & STANDARDS FOR REPORTING OF DIAGNOSTIC, A. 2003a. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Radiology,* 226**,** 24-8.

BOSSUYT, P. M., REITSMA, J. B., BRUNS, D. E., GATSONIS, C. A., GLASZIOU, P. P., IRWIG, L. M., MOHER, D., RENNIE, D., DE VET, H. C., LIJMER, J. G. & STANDARDS FOR REPORTING OF DIAGNOSTIC, A. 2003b. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem,* 49**,** 7-18.

BOSSUYT, P. M. M. 2008b. Interpreting Diagnostic Test Accuracy Studies. *Seminars in Hematology,* 45**,** 189-195.

BOSSUYT, P. M. M. 2009b. Diagnostic accuracy reporting guidelines should prescribe reporting, not modeling. *Journal of clinical epidemiology,* 62**,** 355-356.

BOSSUYT, P. M. M. 2014. STARD (STAndards for Reporting of Diagnostic Accuracy Studies). *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

BOUD, D. & TENNANT, M. 2006. Putting doctoral education to work: challenges to academic practice. *Higher Education Research & Development,* 25**,** 293-306.

BOURSIER, J., DE LEDINGHEN, V., POYNARD, T., GUÉCHOT, J., CARRAT, F., LEROY, V., WONG, G. L.-H., FRIEDRICH-RUST, M., FRAQUELLI, M. & PLEBANI, M. 2015. An extension of STARD statements for reporting diagnostic accuracy studies on liver fibrosis tests: the Liver-FibroSTARD standards. *Journal of hepatology,* 62**,** 807-815.

BRADBURY, H. & REASON, P. 2006. *Handbook of action research*, Sage.

BRAT, R., YOUSEF, N., KLIFA, R., REYNAUD, S., AGUILERA, S. S. & DE LUCA, D. 2015. Lung Ultrasonography Score to Evaluate Oxygenation and Surfactant Need in Neonates Treated With Continuous Positive Airway Pressure. *JAMA pediatrics,* 169**,** e151797-e151797.

BREALEY, S. & SCALLY, A. J. 2008. Methodological approaches to evaluating the practice of radiographers' interpretation of images: A review. *Radiography,* 14**,** e46-e54.

BRELL, M., IBÁÑEZ, J. & TORTOSA, A. 2011. O6-Methylguanine-DNA methyltransferase protein expression by immunohistochemistry in brain and non-brain systemic tumours: systematic review and meta-analysis of correlation with methylation-specific polymerase chain reaction. *BMC cancer,* 11**,** 35.

BRODIE, P. & IRVING, K. 2007. Assessment in work-based learning: investigating a pedagogical approach to enhance student learning. *Assessment & Evaluation in Higher Education,* 32**,** 11-19.

BRUCHER, N., VIAL, J., BAUNIN, C., LABARRE, D., MEYRIGNAC, O., JURICIC, M., BOUALI, O., ABBO, O., GALINIER, P. & SANS, N. 2015. Non-contrast-enhanced MR angiography using time-spin labelling inversion pulse technique for detecting crossing renal vessels in children with symptomatic ureteropelvic junction obstruction: comparison with surgical findings. *European radiology***,** 1-8.

BRYDON-MILLER, M. & MAGUIRE, P. 2009. Participatory action research: Contributions to the development of practitioner inquiry in education. *Educational Action Research,* 17**,** 79-93.

BUDOVEC, J. J. & KAHN, C. E. 2010. Evidence-Based Radiology: A Primer in Reading Scientific Articles. *American Journal of Roentgenology,* 195**,** W1-W4.

BUDOVEC, J. J. & KAHN JR, C. E. 2010. Evidence-based radiology: a primer in reading scientific articles. *American Journal of Roentgenology,* 195**,** W1-W4.

BUNNISS, S. & KELLY, D. R. 2010. Research paradigms in medical education research. *Medical Education,* 44**,** 358-366.

BURCH, J., SOARES-WEISER, K., ST JOHN, D., DUFFY, S., SMITH, S., KLEIJNEN, J. & WESTWOOD, M. 2007. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: a systematic review. *Journal of Medical Screening,* 14**,** 132-137.

BURNS, A. 2005. Action research: An evolving paradigm? *Language teaching,* 38**,** 57-74.

CANTISANI, V., MACERONI, P., D'ANDREA, V., PATRIZI, G., DI SEGNI, M., DE VITO, C., GRAZHDANI, H., ISIDORI, A. M., GIANNETTA, E. & REDLER, A. 2015. Strain ratio ultrasound elastography increases the accuracy of colour-Doppler ultrasound in the evaluation of Thy-3 nodules. A bi-centre university experience. *European radiology***,** 1-9.

CARP, J. 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage,* 63**,** 289.

CASSINOTTO, C., LAPUYADE, B., AÏT-ALI, A., VERGNIOL, J., GAYE, D., FOUCHER, J., BAILACQ-AUDER, C., CHERMAK, F., LE BAIL, B. & DE LÉDINGHEN, V. 2013. Liver fibrosis: noninvasive assessment with acoustic radiation force impulse elastography—comparison with FibroScan M and XL probes and FibroTest in patients with chronic liver disease. *Radiology,* 269**,** 283-292.

CENGIZ, M., SENTÜRK, S., CETIN, B., BAYRAK, A. H. & BILEK, S. U. 2014. Sonographic assessment of fatty liver: intraobserver and interobserver variability. *International Journal of Clinical and Experimental Medicine,* 7**,** 5453-5460.

CHANTEAU, S., DARTEVELLE, S., MAHAMANE, A. E., DJIBO, S., BOISIER, P. & NATO, F. 2006. New rapid diagnostic tests for Neisseria meningitidis serogroups A, W135, C, and Y. *PLoS Med,* 3**,** e337.

CHAVEZ, C. A., SKI, C. F. & THOMPSON, D. R. 2014. Psychometric properties of the Cardiac Depression Scale: A systematic review. *Heart, Lung and Circulation,* 23**,** 610-618.

CHEN, W., MO, J.-J., LIN, L., LI, C.-Q. & ZHANG, J.-F. 2015a. Diagnostic value of magnetic resonance cholangiopancreatography in choledocholithiasis. *World journal of gastroenterology: WJG,* 21**,** 3351.

CHEN, W., XING, W., PENG, Y., HE, Z., WANG, C. & WANG, Q. 2013. Cerebral Aneurysms: Accuracy of 320–Detector Row Nonsubtracted and Subtracted Volumetric CT Angiography for Diagnosis. *Radiology,* 269**,** 841-849.

CHEN, X., YANG, Y., GAN, W., XU, L., YE, Q. & GUO, H. 2015b. Newly Designed Break-Apart and ASPL-TFE3 Dual-Fusion FISH Assay Are Useful in Diagnosing Xp11. 2 Translocation Renal Cell Carcinoma and ASPL-TFE3 Renal Cell Carcinoma: A STARD-Compliant Article. *Medicine,* 94**,** 1-8.

CHIESA, C., PACIFICO, L., NATALE, F., HOFER, N., OSBORN, J. F. & RESCH, B. 2015a. Fetal and early neonatal interleukin-6 response. *Cytokine*.

CHIESA, C., PACIFICO, L., OSBORN, J. F., BONCI, E., HOFER, N. & RESCH, B. 2015b. Early-Onset Neonatal Sepsis: Still Room for Improvement in Procalcitonin Diagnostic Accuracy Studies. *Medicine,* 94.

CHO, N., IM, S.-A., KANG, K. W., PARK, I.-A., SONG, I. C., LEE, K.-H., KIM, T.-Y., LEE, H., CHUN, I. K. & YOON, H.-J. 2015. Early prediction of response to neoadjuvant chemotherapy in breast cancer patients: comparison of single-voxel 1H-magnetic resonance spectroscopy and 18F-fluorodeoxyglucose positron emission tomography. *European radiology***,** 1-12.

CID, J., AGUINACO, R., SÁNCHEZ, R., GARCÍA-PARDO, G. & LLORENTE, A. 2010. Neutrophil CD64 expression as marker of bacterial infection: a systematic review and meta-analysis. *Journal of Infection,* 60**,** 313-319.

CLIBBENS, N., WALTERS, S. & BAIRD, W. 2012. Delphi research: issues raised by a pilot study. *Nurse researcher,* 19**,** 37.

CONNELLY, L. M. 2008. Bland-Altman plots. *Medsurg nursing : official journal of the Academy of Medical-Surgical Nurses,* 17**,** 175.

COOK, T. 2009. The purpose of mess in action research: building rigour though a messy turn. *Educational Action Research,* 17**,** 277-291.

COPPUS, S. F., VAN DER VEEN, F., BOSSUYT, P. M. & MOL, B. W. 2006. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. *Fertility and sterility,* 86**,** 1321-1329.

CORDONNIER, C., SALMAN, R. A.-S. & WARDLAW, J. 2007. Spontaneous brain microbleeds: systematic review, subgroup analyses and standards for study design and reporting. *Brain,* 130**,** 1988-2003.

COSSE, C., SABBAGH, C., KAMEL, S., GALMICHE, A. & REGIMBEAU, J.-M. 2014. Procalcitonin and intestinal ischemia: A review of the literature. *World journal of gastroenterology: WJG,* 20**,** 17773.

COSTA, E. A., CUNHA, G. M., SMORODINSKY, E., CRUITE, I., TANG, A., MARKS, R. M., CLARK, L., WOLFSON, T., GAMST, A. & SICKLICK, J. K. 2015. Diagnostic Accuracy of Preoperative Gadoxetic Acid–enhanced 3-T MR Imaging for Malignant Liver Lesions by Using Ex Vivo MR Imaging–matched Pathologic Findings as the Reference Standard. *Radiology***,** 142069.

COSTLEY, C., ELLIOTT, G. C. & GIBBS, P. 2010. *Doing work based research: Approaches to enquiry for insider-researchers*, Sage.

COSTLEY, C. G., P. 2006. Researching others: care as an ethic for practitioner researchers. *Studies in Higher Education,* 31**,** 89-98.

CRIM, J. R., LAYFIELD, L. J., SCHMIDT, R., HANRAHAN, C., LIU, T. & MANCASTER, B. J. 2013. MRI leads to increased false-positive diagnosis of chondrosarcoma. *Skeletal Radiology,* 42**,** 1044-1045.

DAWSON, G. F. 2012. *Easy Interpretation of Biostatistics: The Vital Link to Applying Evidence in Medical Decisions*, Elsevier Health Sciences.

DUNT, D. & MCKENZIE, R. 2012. Improving the quality of qualitative studies: do reporting guidelines have a place? *Family practice,* 29**,** 367-369.

ELIKASHVILI, I., TAY, E. T. & TSUNG, J. W. 2014. The Effect of Point-of-care Ultrasonography on Emergency Department Length of Stay and Computed Tomography Utilization in Children With Suspected Appendicitis. *Academic Emergency Medicine,* 21**,** 163-170.

ENG, J. 2003. Sample Size Estimation: How Many Individuals Should Be Studied?1. *Radiology,* 227**,** 309-313.

ENG, J. 2004. Sample Size Estimation: A Glimpse beyond Simple Formulas 1. *Radiology,* 230**,** 606-612.

ERAUT, M. 2000. Non-formal learning and tacit knowledge in professional work. *British journal of educational psychology,* 70**,** 113-136.

ERRICO, G., GIORDANO, A. & PALTRINIERI, S. 2012. Diagnostic accuracy of electrophoretic analysis of native or defribrinated plasma using serum as a reference sample. *Veterinary Clinical Pathology,* 41**,** 529-540.

ETHERINGTON, K. 2004. *Becoming a reflexive researcher: Using our selves in research*, Jessica Kingsley Publishers.

FALLENBERG, E., DROMAIN, C., DIEKMANN, F., ENGELKEN, F., KROHN, M., SINGH, J., INGOLD-HEPPNER, B., WINZER, K., BICK, U. & RENZ, D. 2014. Contrast-enhanced spectral mammography versus MRI: initial results in the detection of breast cancer and assessment of tumour size. *European radiology,* 24**,** 256-264.

FERREIRA, A. & PACHECO, A. 2015. SimTCM: A human patient simulator with application to diagnostic accuracy studies of Chinese medicine. J Integr Med. 2015; 13 (1): 9–19.

FIDALGO, B. M., CRABB, D. P. & LAWRENSON, J. G. 2015. Methodology and reporting of diagnostic accuracy studies of automated perimetry in glaucoma: evaluation using a standardised approach. *Ophthalmic and Physiological Optics,* 35**,** 315-323.

FLANIGAN, T. S., MCFARLANE, E. & COOK, S. Conducting survey research among physicians and other medical professionals: A review of current literature. Proceedings of the Survey Research Methods Section, American Statistical Association, 2008. 4136-47.

FLETCHER, A. J. & MARCHILDON, G. P. 2014. Using the delphi method for qualitative, participatory action research in health leadership. *International Journal of Qualitative Methods,* 13**,** 1-18.

FLICKER, L., RITCHIE, C., NOEL-STORR, A. & MCSHANE, R. 2012. Harmonization of reporting standards for studies of diagnostic test accuracy in dementia and related conditions: the STARDdem (STAndards for the Reporting of Diagnostic accuracy studies-Dementia) criteria. *Alzheimer's & Dementia,* 8**,** P106.

FLORKOWSKI, C. M. 2008. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews,* 29**,** S83.

FONTELA, P. S., PANT PAI, N., SCHILLER, I., DENDUKURI, N., RAMSAY, A. & PAI, M. 2009. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PLoS One,* 4**,** e7753.

FOX, M., GREEN, G. & MARTIN, P. 2007. *Doing practitioner research*, Sage.

FRATZ, S., SCHUHBAECK, A., BUCHNER, C., BUSCH, R., MEIERHOFER, C., MARTINOFF, S., HESS, J. & STERN, H. 2009. Comparison of accuracy of axial slices versus short-axis slices for measuring ventricular volumes by cardiac magnetic resonance in patients with corrected tetralogy of fallot. *The American journal of cardiology,* 103**,** 1764-1769.

FREEMAN, K., SZCZEPURA, A. & OSIPENKO, L. 2009. Non-invasive fetal RHD genotyping tests: a systematic review of the quality of reporting of diagnostic accuracy in published studies. *European Journal of Obstetrics & Gynecology and Reproductive Biology,* 142**,** 91-98.

GARDNER, I. A., BURNLEY, T. & CARAGUEL, C. 2014. Improvements are Needed in Reporting of Accuracy Studies for Diagnostic Tests Used for Detection of Finfish Pathogens. *Journal of aquatic animal health,* 26**,** 203-209.

GATSONIS, C. 2003. Do we need a checklist for reporting the results of diagnostic test evaluations? The STARD proposal. *Acad Radiol,* 10**,** 599-600.

GEEVASINGA, N., MENON, P., YIANNIKAS, C., HOWELLS, J., KIERNAN, M. & VUCIC, S. 2015. Threshold tracking TMS: A novel diagnostic technique for Amyotrophic Lateral Sclerosis (S24. 005). *Neurology,* 84**,** S24. 005.

GEFFROY, Y., BOULAY-COLETTA, I., JULLÈS, M.-C., NAKACHE, S., TAOUREL, P. & ZINS, M. 2014. Increased unenhanced bowel-wall attenuation at multidetector CT is highly specific of ischemia complicating small-bowel obstruction. *Radiology,* 270**,** 159-167.

GEORGANTOPOULOU, C., SIMM, A. & ROBERTS, M. 2008. Transvaginal saline hysterosonography: a comparison with local anaesthetic hysteroscopy for the diagnosis of benign lesions associated with menorrhagia. *Gynecological Surgery,* 5**,** 27-34.

GLASZIOU, P., ALTMAN, D. G., BOSSUYT, P., BOUTRON, I., CLARKE, M., JULIOUS, S., MICHIE, S., MOHER, D. & WAGER, E. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet,* 383**,** 267-276.

GODLEE, F. 1994. The Cochrane collaboration. *BMJ : British Medical Journal,* 309**,** 969-970.

GOEBELL, P. J., KAMAT, A. M., SYLVESTER, R. J., BLACK, P., DROLLER, M., GODOY, G., M'LISS, A. H., JUNKER, K., KASSOUF, W. & KNOWLES, M. A. Assessing the quality of studies on the diagnostic accuracy of tumor markers. Urologic Oncology: Seminars and Original Investigations, 2014. Elsevier, 1051-1060.

GRANT, J. 2002. Learning needs assessment: assessing the need. *Bmj,* 324**,** 156-159.

GRAY, D. E. 2013. *Doing research in the real world*, Sage.

GRIFFITHS, P., JARVIS, D., MCQUILLAN, H., WILLIAMS, F., PALEY, M. & ARMITAGE, P. 2013. MRI of the foetal brain using a rapid 3D steady-state sequence. *The British journal of radiology,* 86**,** 20130168.

GRIX, J. 2004. The foundations of research. Basingstoke; New York: Palgrave Macmillan.

GRIX, J. 2010. *The foundations of research*, Palgrave Macmillan.

GUBA, E. G. & LINCOLN, Y. S. 1994. Competing paradigms in qualitative research. *Handbook of qualitative research,* 2.

GUSTAVSEN, B. 2001. Theory and practice: The mediating discourse. *Handbook of action research: The concise paperback edition***,** 17-26.

GUYATT, G., CAIRNS, J., CHURCHILL, D., COOK, D., HAYNES, B., HIRSH, J., IRVINE, J., LEVINE, M., LEVINE, M. & NISHIKAWA, J. 1992. Evidence-based medicine: a new approach to teaching the practice of medicine. *Jama,* 268**,** 2420-2425.

H-ICI, D. O., RIDGWAY, J. P., KUEHNE, T., BERGER, F., PLEIN, S., SIVANANTHAN, M. & MESSROGHLI, D. R. 2012. Cardiovascular magnetic resonance of myocardial edema using a short inversion time inversion recovery (STIR) black-blood technique: Diagnostic accuracy of visual and semi-quantitative assessment. *Journal of Cardiovascular Magnetic Resonance,* 14**,** 22-22.

HADDOW, L. J., FLOYD, S., COPAS, A. & GILSON, R. 2013. A systematic review of the screening accuracy of the HIV Dementia Scale and International HIV Dementia Scale. *PloS one,* 8**,** e61826.

HÅKONSEN, S. J., PEDERSEN, P. U., BATH-HEXTALL, F. & KIRKPATRICK, P. 2015. Diagnostic test accuracy of nutritional tools used to identify undernutrition in patients with colorectal cancer: a systematic review. *The JBI Database of Systematic Reviews and Implementation Reports,* 13**,** 141-187.

HALL, S., LEWITH, G., BRIEN, S. & LITTLE, P. 2008. A review of the literature in applied and specialised kinesiology. *Forschende Komplementärmedizin/Research in Complementary Medicine,* 15**,** 40-46.

HARRISON, J. K., FEARON, P., NOEL-STORR, A. H., MCSHANE, R., STOTT, D. J. & QUINN, T. J. 2014. Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within a general practice (primary care) setting. *Cochrane Database of Systematic Reviews,* 7.

HAUTH, E., HOHMUTH, H., COZUB-POETICA, C., BERNAND, S., BEER, M. & JAEGER, H. 2015. Multiparametric MRI of the prostate with three functional techniques in patients with PSA elevation before initial TRUS-guided biopsy. *The British journal of radiology,* 88**,** 20150422.

HAYNES, R. B. & WILCZYNSKI, N. L. 2004. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. *Bmj,* 328**,** 1040.

HELLEMONS, M. E., KERSCHBAUM, J., BAKKER, S. J., NEUWIRT, H., MAYER, B., MAYER, G., DE ZEEUW, D., LAMBERS HEERSPINK, H. & RUDNICKI, M. 2012. Validity of biomarkers predicting onset or progression of nephropathy in patients with Type 2 diabetes: a systematic review. *Diabetic Medicine,* 29**,** 567-577.

HENES, F., NÜCHTERN, J., GROTH, M., HABERMANN, C., REGIER, M., RUEGER, J., ADAM, G. & GROSSTERLINDEN, L. 2012. Comparison of diagnostic accuracy of magnetic resonance imaging and multidetector computed tomography in the detection of pelvic fractures. *European journal of radiology,* 81**,** 2337-2342.

HERON, J. & REASON, P. 1997. A participatory inquiry paradigm. *Qualitative inquiry,* 3**,** 274-294.

HERR, K. & ANDERSON, G. L. 2005. *The action research dissertation: A guide for students and faculty*, Sage.

HEWITT, M., MCPHAIL, M., POSSAMAI, L., VLAVIANOS, P., DHAR, A. & MONAHAN, K. 2011. A meta-analysis of endoscopic ultrasound with fine needle aspiration (EUS-FNA) for diagnosis of solid pancreatic neoplasms. *Gut,* 60**,** A190-A191.

HING, W., WHITE, S., REID, D. & MARSHALL, R. 2009. Validity of the McMurray's test and modified versions of the test: a systematic literature review. *Journal of Manual & Manipulative Therapy,* 17**,** 22-35.

HIRAMITSU, T., TOMINAGA, Y., OKADA, M., YAMAMOTO, T. & KOBAYASHI, T. 2015. A Retrospective Study of the Impact of Intraoperative Intact Parathyroid Hormone Monitoring During Total Parathyroidectomy for Secondary Hyperparathyroidism: STARD Study. *Medicine,* 94**,** e1213.

HSU, C.-C. & SANDFORD, B. A. 2007. The Delphi technique: making sense of consensus. *Practical assessment, research & evaluation,* 12**,** 1-8.

HUDON, C., FORTIN, M., HAGGERTY, J. L., LAMBERT, M. & POITRAS, M.-E. 2011. Measuring patients' perceptions of patient-centered care: a systematic review of tools for family medicine. *The Annals of Family Medicine,* 9**,** 155-164.

INSTITUTE, S. 2012. *SAS/STAT 12.1 User's Guide Survival Analysis (book Excerpt)*, SAS Institute Incorporated.

JAHROMI, A. S., CINÀ, C. S., LIU, Y. & CLASE, C. M. 2005. Sensitivity and specificity of color duplex ultrasound measurement in the estimation of internal carotid artery stenosis: a systematic review and meta-analysis. *Journal of vascular surgery,* 41**,** 962-972.

JONES, C. M. & ATHANASIOU, T. 2009. Diagnostic accuracy meta-analysis: review of an important tool in radiological research and decision making. *Br J Radiol,* 82**,** 441-6.

KEMMIS, S. & MCTAGGART, R. 2005. Communicative action and the public sphere. *Denzin, NK & Lincoln, YS (red.), The Sage handbook of qualitative research,* 3**,** 559-603.

KIM, W. H., CHANG, J. M., MOON, H.-G., YI, A., KOO, H. R., GWEON, H. M. & MOON, W. K. 2015. Comparison of the diagnostic performance of digital breast tomosynthesis and magnetic resonance imaging added to digital mammography in women with known breast cancers. *European radiology***,** 1-9.

KINDON, S. & ELWOOD, S. 2009. Introduction: More than Methods—Reflections on Participatory Action Research in Geographic Teaching, Learning and Research: Participatory Action Research in Geographic Teaching, Learning and Research. *Journal of Geography in Higher Education,* 33**,** 19-32.

KNOTTNERUS, J. A. & MURIS, J. W. 2003. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *Journal of clinical epidemiology,* 56**,** 1118-1128.

KOH, K.-J., LIST, T., PETERSSON, A. & ROHLIN, M. 2008. Relationship between clinical and magnetic resonance imaging diagnoses and findings in degenerative and inflammatory temporomandibular joint diseases: a systematic literature review. *Journal of orofacial pain,* 23**,** 123-139.

KOREVAAR, D. A., WANG, J., VAN ENST, W. A., LEEFLANG, M. M., HOOFT, L., SMIDT, N. & BOSSUYT, P. M. 2014. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. *Radiology,* 274**,** 781-789.

KOSHY, E., KOSHY, V. & WATERMAN, H. 2011. Action Research in Healthcare. London: Sage.

KUNATH, F., GROBE, H. R., RÜCKER, G., ENGEHAUSEN, D., ANTES, G., WULLICH, B. & MEERPOHL, J. J. 2012. Do journals publishing in the field of urology endorse reporting guidelines? A survey of author instructions. *Urologia internationalis,* 88**,** 54-59.

LAMÉRIS, W., VAN RANDEN, A., VAN ES, H. W., VAN HEESEWIJK, J. P. M., VAN RAMSHORST, B., BOUMA, W. H., TEN HOVE, W., VAN LEEUWEN, M. S., VAN KEULEN, E. M., DIJKGRAAF, M. G. W., BOSSUYT, P. M. M., BOERMEESTER, M. A. & STOKER, J. 2009. *Imaging strategies for detection of urgent conditions in patients with acute abdominal pain: diagnostic accuracy study*.

LANGLOIS, S., GOUDREAU, J. & LALONDE, L. 2014. Scientific rigour and innovations in participatory action research investigating workplace learning in continuing interprofessional education. *Journal of Interprofessional Care,* 28**,** 226-231.

LAWRENCE, R. J. & DESPRÉS, C. 2004. Futures of transdisciplinarity. *Futures,* 36**,** 397-405.

LEEFLANG, M. M. 2015. Reporting diagnostic accuracy studies: where are we now? *Biomarkers in medicine,* 9**,** 897-899.

LEEFLANG, M. M. G., BOSSUYT, P. M. M. & IRWIG, L. 2009. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *Journal of clinical epidemiology,* 62**,** 5-12.

LEES, R., SELVARAJAH, J., FENTON, C., PENDLEBURY, S. T., LANGHORNE, P., STOTT, D. J. & QUINN, T. J. 2014. Test accuracy of cognitive screening tests for diagnosis of dementia and multidomain cognitive impairment in stroke. *Stroke,* 45**,** 3008-3018.

LENTLE, B. C., ARNOLD, R. E., BECKER, G. J., BRYAN, R. N., FRITZSCHE, P. J. & HUSSEY, D. H. 2007. What We Do Not Yet Know in the Radiologic Sciences 1. *Radiology,* 243**,** 618-621.

LIGOCKI, C., ABADEH, A., WANG, K. C., ADAMS-WEBBER, T. & DORIA, A. 2015. A systematic review of ultrasound imaging as a tool for evaluating hemophilic arthropathy in children. *Pediatric Radiology,* 45**,** S234.

LIJMER, J. G., MOL, B. W., HEISTERKAMP, S., BONSEL, G. J., PRINS, M. H., VAN DER MEULEN, J. H. & BOSSUYT, P. M. 1999. Empirical evidence of design-related bias in studies of diagnostic tests. *Jama,* 282**,** 1061-1066.

LUMBRERAS, B., JARRÍN, I. & HERNÁNDEZ AGUADO, I. 2006. Evaluation of the research methodology in genetic, molecular and proteomic tests. *Gaceta Sanitaria,* 20**,** 368-373.

LUNDH, A. & GØTZSCHE, P. C. 2008. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC medical research methodology,* 8**,** 22.

MACKENZIE, N. & KNIPE, S. 2006. Research dilemmas: Paradigms, methods and methodology. *Issues in educational research,* 16**,** 193-205.

MACLEAN, E. N., STONE, I. S., CEELEN, F., GARCIA-ALBENIZ, X., SOMMER, W. H. & PETERSEN, S. E. 2014. Reporting standards in cardiac MRI, CT, and SPECT diagnostic accuracy studies: analysis of the impact of STARD criteria. *European Heart Journal-Cardiovascular Imaging***,** jet277.

MAHONEY, J. & ELLISON, J. 2007. Assessing the quality of glucose monitor studies: a critical evaluation of published reports. *Clinical chemistry,* 53**,** 1122-1128.

MALCIUS, D., JONKUS, M., KUPRIONIS, G., MALECKAS, A., MONASTYRECKIENE, E., UKTVERIS, R., RINKEVICIUS, S. & BARAUSKAS, V. 2009. The accuracy of different imaging techniques in diagnosis of acute hematogenous osteomyelitis. *Medicina (Kaunas),* 45**,** 624-631.

MANCHIKANTI, L., DERBY, R., WOLFER, L., SINGH, V., DATTA, S. & HIRSCH, J. A. 2009. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 5. Diagnostic accuracy studies. *Pain Physician,* 12**,** 517-40.

MARTIN, J., WILLIAMS, K., SUTTON, A. J., ABRAMS, K. R. & ASSASSA, R. 2006. Systematic review and meta-analysis of methods of diagnostic assessment for urinary incontinence. *Neurourology and urodynamics,* 25**,** 674-683.

MCCOMISKEY, M. H., MCCLUGGAGE, W. G., GREY, A., HARLEY, I., DOBBS, S. & NAGAR, H. A. 2012. Diagnostic accuracy of magnetic resonance imaging in endometrial cancer. *International Journal of Gynecological Cancer,* 22**,** 1020-1025.

MCNAMEE, L. S., O'BRIEN, F. Y. & BOTHA, J. H. 2009. Student perceptions of medico-legal autopsy demonstrations in a student-centred curriculum. *Medical education,* 43**,** 66-73.

MCNIFF, J. 1995. *Action research for professional development*, Hyde Publications Bournemouth.

MCNIFF, J. 2002. Action research for professional development Concise advice for new action researchers.

MCNIFF, J. 2013. *Action research: Principles and practice*, Routledge.

MCNIFF, J. & WHITEHEAD, A. 2002. Action Research: Principles and Practice.

MEDINA, L. S. & BLACKMORE, C. C. 2007. Evidence-based Radiology: Review and Dissemination 1. *Radiology,* 244**,** 331-336.

MEDINA, L. S. & ZURAKOWSKI, D. 2003. Measurement Variability and Confidence Intervals in Medicine: Why Should Radiologists Care?1. *Radiology,* 226**,** 297-301.

MENON, P., GEEVASINGA, N., YIANNIKAS, C., HOWELLS, J., KIERNAN, M. C. & VUCIC, S. 2015. Sensitivity and specificity of threshold tracking transcranial magnetic stimulation for diagnosis of amyotrophic lateral sclerosis: a prospective study. *The Lancet Neurology,* 14**,** 478-484.

MEYER, J. 2000. Qualitative research in health care: Using qualitative methods in health related action research. *BMJ: British Medical Journal,* 320**,** 178.

MIERITZ, R. M., BRONFORT, G., KAWCHUK, G., BREEN, A. & HARTVIGSEN, J. 2012. Reliability and measurement error of 3-dimensional regional lumbar motion measures: a systematic review. *Journal of manipulative and physiological therapeutics,* 35**,** 645-656.

MILLER, E., ROPOSCH, A., ULERYK, E. & DORIA, A. S. 2009. Juvenile Idiopathic Arthritis of Peripheral Joints: Quality of Reporting of Diagnostic Accuracy of Conventional MRI1. *Academic radiology,* 16**,** 739-757.

MITCHELL, A. J. & COYNE, J. C. 2007. Do ultra-short screening instruments accurately detect depression in primary care? *British Journal of General Practice,* 57**,** 144-151.

MOHER, D., ALTMAN, D., SCHULZ, K., SIMERA, I. & WAGER, E. 2014a. *Guidelines for reporting health research: a user's manual*, Wiley Online Library.

MOHER, D., ALTMAN, D. G., SCHULZ, K. F. & SIMERA, I. 2014b. How to Develop a Reporting Guideline. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

MOHER, D., PLINT, A. C., ALTMAN, D. G., SCHULZ, K. F., KOBER, T., GALLOWAY, E. K., WEEKS, L. & DIAS, S. 2010a. Consolidated standards of reporting trials (CONSORT) and the quality of reporting of randomized controlled trials. *The Cochrane Library*.

MOHER, D., SCHULZ, K. F., ALTMAN, D. G., HOEY, J., GRIMSHAW, J., MILLER, D., SEELY, D., SIMERA, I., SAMPSON, M., WEEKS, L. & OCAMPO, M. 2014c. Characteristics of Available Reporting Guidelines. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

MOHER, D., SCHULZ, K. F., SIMERA, I. & ALTMAN, D. G. 2010b. Guidance for developers of health research reporting guidelines. *PLoS medicine,* 7**,** e1000217.

MOHER, D., WEEKS, L., OCAMPO, M., SEELY, D., SAMPSON, M., ALTMAN, D. G., SCHULZ, K. F., MILLER, D., SIMERA, I. & GRIMSHAW, J. 2011. Describing reporting guidelines for health research: a systematic review. *Journal of Clinical Epidemiology,* 64**,** 718-742.

MONTORI, V. M., JAESCHKE, R., SCHÜNEMANN, H. J., BHANDARI, M., BROZEK, J. L., DEVEREAUX, P. J. & GUYATT, G. H. 2004. Users' Guide To Detecting Misleading Claims In Clinical Research Reports. *BMJ: British Medical Journal,* 329**,** 1093-1096.

MORGAN, D. L. 2007. Paradigms lost and pragmatism regained methodological implications of combining qualitative and quantitative methods. *Journal of mixed methods research,* 1**,** 48-76.

MORRISON, B. & LILFORD, R. 2001. How can action research apply to health services? *Qualitative Health Research,* 11**,** 436-449.

MYBURGH, C., LARSEN, A. H. & HARTVIGSEN, J. 2008. A systematic, critical review of manual palpation for identifying myofascial trigger points: evidence and clinical significance. *Archives of physical medicine and rehabilitation,* 89**,** 1169-1176.

NAGAR, H., DOBBS, S., MCCLELLAND, H. R., PRICE, J., MCCLUGGAGE, W. G. & GREY, A. 2006. The diagnostic accuracy of magnetic resonance imaging in detecting cervical involvement in endometrial cancer. *Gynecologic oncology,* 103**,** 431-434.

NETWORK, E. 2009. Enhancing the quality and transparency of health research. *Available at www. equator-network. org*.

NIEUWENHUIJZE, M. J., KORSTJENS, I., DE JONGE, A., DE VRIES, R. & LAGRO-JANSSEN, A. 2014. On speaking terms: a Delphi study on shared decision-making in maternity care. *BMC Pregnancy and Childbirth,* 14**,** 223-223.

NOEL-STORR, A. H., FLICKER, L., RITCHIE, C. W., NGUYEN, G. H., GUPTA, T., WOOD, P., WALTON, J., DESAI, M., SOLOMON, D. F. & MOLENA, E. 2013. Systematic review of the body of evidence for the use of biomarkers in the diagnosis of dementia. *Alzheimer's & Dementia,* 9**,** e96-e105.

O'LEARY, J. D. & CRAWFORD, M. W. 2013. Review article: Reporting guidelines in the biomedical literature. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie,* 60**,** 813-821.

OBUCHOWSKI, N. A. 2003a. Receiver operating characteristic curves and their use in radiology. *Radiology,* 229**,** 3-8.

OBUCHOWSKI, N. A. 2003b. Special Topics III: bias. *Radiology,* 229**,** 617-621.

OCHODO, E. A., DE HAAN, M. C., REITSMA, J. B., HOOFT, L., BOSSUYT, P. M. & LEEFLANG, M. M. 2013. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin". *Radiology,* 267**,** 581-8.

PARK, P. 2006. Knowledge and participatory research. *Handbook of action research: Concise paperback edition***,** 83-93.

PARRY, R. A., GLAZE, S. A. & ARCHER, B. R. 1999. The AAPM/RSNA physics tutorial for residents. Typical patient radiation doses in diagnostic radiology. *Radiographics : a review publication of the Radiological Society of North America, Inc,* 19**,** 1289.

PAVLOV, C., CASAZZA, G., NIKOLOVA, D., TSOCHATZIS, E., BURROUGHS, A., IVASHKIN, V. & GLUUD, C. 2015. Transient elastography for diagnosis of stages of hepatic fibrosis and cirrhosis in people with alcoholic liver disease. *status and date: New, published in*.

PERRY, A. E., MARANDOS, R., COULTON, S. & JOHNSON, M. 2010. Screening tools assessing risk of suicide and self-harm in adult offenders: a systematic review. *International journal of offender therapy and comparative criminology*.

PLOUS, S. 1993. *The psychology of judgment and decision making*, Mcgraw-Hill Book Company.

POLDRACK, R. A., FLETCHER, P. C., HENSON, R. N., WORSLEY, K. J., BRETT, M. & NICHOLS, T. E. 2008. Guidelines for reporting an fMRI study. *NeuroImage,* 40**,** 409-414.

POPE, A. 2009. Reproducibility: Intraobserver and Interobserver Variability.

PRUMMEL, M. V., MURADALI, D., SHUMAK, R., MAJPRUZ, V., BROWN, P., JIANG, H., DONE, S. J., YAFFE, M. J. & CHIARELLI, A. M. 2015. Digital Compared with Screen-Film Mammography: Measures of Diagnostic Accuracy among Women Screened in the Ontario Breast Screening Program. *Radiology***,** 150733.

RAHMAN, R. L., CRAWFORD, S. L. & SIWAWA, P. 2015. Management of axilla in breast cancer–The saga continues. *The Breast*.

RAJA, A. S., PINES, J. M., SCHUUR, J. D., MUIR, M., CALFEE, R. P. & CARPENTER, C. R. 2013. Evidence based diagnostics: Meta-analysis of the accuracy of physical exam and imaging for adult scaphoid fractures. *Academic Emergency Medicine,* 20**,** S24-S25.

RAMA, K. R. B. S., POOVALI, S. & APSINGI, S. 2006. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clinical orthopaedics and related research,* 447**,** 237-246.

RAUTIAINEN, S., MASARWAH, A., SUDAH, M., SUTELA, A., PELKONEN, O., JOUKAINEN, S., SIRONEN, R., KÄRJÄ, V. & VANNINEN, R. 2013. Axillary lymph node biopsy in newly diagnosed invasive breast cancer: comparative accuracy of fine-needle aspiration biopsy versus core-needle biopsy. *Radiology,* 269**,** 54-60.

RCR. 2012. *Good practice guide for clinical radiologists* [Online]. London. Available: https://www.rcr.ac.uk/sites/default/files/publication/BFCR%2812%291_GoodPractice.pdf.

REASON, P. & BRADBURY, H. 2001. Introduction: Inquiry and Participation in Search of a World Worthy of Human Aspiration, w: P. Reason, H. Bradbury (red.). *Handbook of Action Research Participative Inquiry and Practice*.

REASON, P. & BRADBURY, H. 2005. *Handbook of action research: concise paperback edition*, Sage.

REITSMA, J. B., RUTJES, A. W. S., KHAN, K. S., COOMARASAMY, A. & BOSSUYT, P. M. 2009. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of clinical epidemiology,* 62**,** 797-806.

RENNIE, D. 2014. Frontmatter. *Guidelines for Reporting Health Research: A User's Manual.* John Wiley & Sons, Ltd.

RIEGELMAN, R. K. 2000. *Studying a study and testing a test: how to read the medical evidence,* Philadelphia, Lippincott Williams & Wilkins.

RIEGELMAN, R. K. 2013. Testing a Test - M.A.A.R.I.E. Framework: Method, Assignment, and Assessment

In: RHYNER, S. (ed.) *Studying A Study & Testing A Test.* Baltimore, MD Lippincott Williams & Wilkins

ROBSON, C. 2002. *Real world research: A resource for social scientists and practitioner-researchers,* Blackwell Oxford.

ROPOSCH, A., MOREAU, N. M., ULERYK, E. & DORIA, A. S. 2006. Developmental dysplasia of the hip: quality of reporting of diagnostic accuracy for US. *Radiology,* 241**,** 854-860.

RUTJES, A. W. S., REITSMA, J. B., DI NISIO, M., SMIDT, N., VAN RIJN, J. C. & BOSSUYT, P. M. M. 2006. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne,* 174**,** 469-476.

RYCHETNIK, L., FROMMER, M., HAWE, P. & SHIELL, A. 2002. Criteria for evaluating evidence on public health interventions. *Journal of epidemiology and community health,* 56**,** 119-127.

SABA, L., GUERRIERO, S., SULCIS, R., PILLONI, M., AJOSSA, S., MELIS, G. & MALLARINI, G. 2012. MRI and "Tenderness Guided" transvaginal ultrasonography in the diagnosis of recto-sigmoid endometriosis. *Journal of Magnetic Resonance Imaging,* 35**,** 352-360.

SABA, L., GUERRIERO, S., SULIS, R., PILLONI, M., AJOSSA, S., MELIS, G. & MALLARINI, G. 2011. Learning curve in the detection of ovarian and deep endometriosis by using Magnetic Resonance: comparison with surgical results. *European journal of radiology,* 79**,** 237-244.

SANDREY, M. A. 2013. Special physical examination tests for superior labrum anterior-posterior shoulder tears: an examination of clinical usefulness. *Journal of athletic training,* 48**,** 856.

SARDANELLI, F. & DI LEO, G. 2009a. *Biostatistics for radiologists: planning, performing, and writing a radiologic study*, Springer Science & Business Media.

SARDANELLI, F. & DI LEO, G. 2009b. Reproducibility: Intraobserver and Interobserver Variability. *Biostatistics for Radiologists.* Springer Milan.

SARDANELLI, F., HUNINK, M. G., GILBERT, F. J., DI LEO, G. & KRESTIN, G. P. 2010. Evidence-based radiology: why and how? *European radiology,* 20**,** 1-15.

SCHATZKI, T. R., KNORR-CETINA, K. D. & SAVIGNY, E. V. (eds.) 2000. *The Practice Turn in Contemporary Theory*: Taylor & Francis Ltd - M.U.A.

SCHÖN, D. 1987a. *Educating the Reflective Practitioner,* San Francisco, Jossey-Bass.

SCHÖN, D. 1987b. Educating the reflective practitioner.

SCHUBERT, T., TAKES, M., ASCHWANDEN, M., KLARHOEFER, M., HAAS, T., JACOB, A. L., LIU, D., GUTZEIT, A. & KOS, S. 2015. Non-enhanced, ECG-gated MR angiography of the pedal vasculature: comparison with contrast-enhanced MR angiography and digital subtraction angiography in peripheral arterial occlusive disease. *European radiology***,** 1-9.

SCOTT, D., BROWN, A. & LUNT, I. 2004. *Professional Doctorates: Integrating Academic And Professional Knowledge: Integrating Academic and Professional Knowledge*, McGraw-Hill Education (UK).

SEITH, F., GATIDIS, S., SCHMIDT, H., BEZRUKOV, I., LA FOUGÈRE, C., NIKOLAOU, K., PFANNENBERG, C. & SCHWENZER, N. 2015. Comparison of Positron Emission Tomography Quantification Using Magnetic Resonance-and Computed Tomography-Based Attenuation Correction in Physiological Tissues and Lesions: A Whole-Body Positron Emission Tomography/Magnetic Resonance Study in 66 Patients. *Investigative radiology*.

SELMAN, T., KHAN, K. & MANN, C. 2005. An evidence-based approach to test accuracy studies in gynecologic oncology: the 'STARD'checklist. *Gynecologic oncology,* 96**,** 575-578.

SELMAN, T. J., MORRIS, R. K., ZAMORA, J. & KHAN, K. S. 2011. The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. *BMC women's health,* 11**,** 8.

SHARMA, S., CROW, H. C., MCCALL JR, W. & GONZALEZ, Y. M. 2012. Systematic review of reliability and diagnostic validity of joint vibration analysis for diagnosis of temporomandibular disorders. *Journal of orofacial pain,* 27**,** 51-60.

SHIVKUMAR, S., PEELING, R., JAFARI, Y., JOSEPH, L. & PAI, N. P. 2012. Accuracy of rapid and point-of-care screening tests for hepatitis C: a systematic review and meta-analysis. *Annals of internal medicine,* 157**,** 558-566.

SHUNMUGAM, M. & AZUARA-BLANCO, A. 2006. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Investigative ophthalmology & visual science,* 47**,** 2317-2323.

SIDDIQUI, I. A., SABAH, S. A., SATCHITHANANDA, K., LIM, A. K., CRO, S., HENCKEL, J., SKINNER, J. A. & HART, A. J. 2014. A comparison of the diagnostic accuracy of MARS MRI and ultrasound of the painful metal-on-metal hip arthroplasty. *Acta Orthopaedica,* 85**,** 375-382.

SIDDIQUI, M., AZUARA-BLANCO, A. & BURR, J. 2005. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *British journal of ophthalmology,* 89**,** 261-265.

SIMERA, I., ALTMAN, D. G., MOHER, D., SCHULZ, K. F. & HOEY, J. 2008. Guidelines for reporting health research: the EQUATOR network's survey of guideline authors. *PLoS Med,* 5**,** e139.

SMIDT, N., ANNE, W. S. R., DANIÃ«LLE, A. W. M. V. D. W., RAYMOND, W. J. G. O., REITSMA, J. B., BOSSUYT, P. M., BOUTER, L. M. & HENRICA, C. W. D. V. 2005. Quality of Reporting of Diagnostic Accuracy Studies1. *Radiology,* 235**,** 347.

SMIDT, N., RUTJES, A., VAN DER WINDT, D., OSTELO, R., BOSSUYT, P., REITSMA, J., BOUTER, L. & DE VET, H. 2006. The quality of diagnostic accuracy studies since the STARD statement Has it improved? *Neurology,* 67**,** 792-797.

SMITH, L., BRATINI, L., CHAMBERS, D.-A., JENSEN, R. V. & ROMERO, L. 2010. Between idealism and reality: Meeting the challenges of participatory action research. *Action Research,* 8**,** 407-425.

SMITH, M. K. 2001. Chris Argyris: theories of action, double-loop learning and organizational learning. *The encyclopedia of informal education,* 1.

SOLOMON, M. 2011. Just a paradigm: evidence-based medicine in epistemological context. *European Journal for Philosophy of Science,* 1**,** 451-466.

SOMEKH, B. 1995. The contribution of action research to development in social endeavours: a position paper on action research methodology. *British Educational Research Journal,* 21**,** 339-355.

SOUSA, V. E. C., LOPES, M. V. D. O. & SILVA, V. M. 2015. Systematic review and meta-analysis of the accuracy of clinical indicators for ineffective airway clearance. *Journal of advanced nursing,* 71**,** 498-513.

STENGEL, D., BAUWENS, K., RADEMACHER, G., MUTZE, S. & EKKERNKAMP, A. 2005. Association between Compliance with Methodological Standards of Diagnostic Research and Reported Test Accuracy: Meta-Analysis of Focused Assessment of US for Trauma 1. *Radiology,* 236**,** 102-111.

STEVENS, A., SHAMSEER, L., WEINSTEIN, E., YAZDI, F., TURNER, L., THIELMAN, J., ALTMAN, D. G., HIRST, A., HOEY, J. & PALEPU, A. 2014. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ: British Medical Journal,* 348.

STOKOLS, D. 2006. Toward a science of transdisciplinary action research. *American journal of community psychology,* 38**,** 63-77.

STRASSLE, P., HESS, A. S., THOM, K. A. & HARRIS, A. D. 2012. Assessing Sensitivity and Specificity in New Diagnostic Tests: The Importance and Challenges of Study Populations. *Infection Control and Hospital Epidemiology,* 33**,** 1177-1178.

STREINER, D. L. 2003. Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of personality assessment,* 81**,** 209-219.

SU, S.-B., QIN, S.-Y., CHEN, W., LUO, W. & JIANG, H.-X. 2015. Carbohydrate antigen 19-9 for differential diagnosis of pancreatic carcinoma and chronic pancreatitis. *World journal of gastroenterology: WJG,* 21**,** 4323.

SU, S.-B., QIN, S.-Y., GUO, X.-Y., LUO, W. & JIANG, H.-X. 2013. Assessment by meta-analysis of interferon-gamma for the diagnosis of tuberculous peritonitis. *World journal of gastroenterology: WJG,* 19**,** 1645.

THAWATCHAI LEELAHANAJ, M. 2010. Developing Thai economic model to study cost-effectiveness of switching to bupropion compared to combination with bupropion after the failure of an SSRI for major depressive disorder. *J Med Assoc Thai,* 93**,** S35-S42.

THIEME, M. E., LEEUWENBURGH, M. M., VALDEHUEZA, Z. D., BOUMAN, D. E., DE BRUIN, I. G., SCHREURS, W. H., HOUDIJK, A. P., STOKER, J. & WIARDA, B. M.

2014. Diagnostic accuracy and patient acceptance of MRI in children with suspected appendicitis. *European radiology,* 24**,** 630-637.

THORNTON, G., MCPHAIL, M., NAYAGAM, S., HEWITT, M., VLAVIANOS, P. & MONAHAN, K. 2013. Endoscopic ultrasound guided fine needle aspiration for the diagnosis of pancreatic cystic neoplasms: a meta-analysis. *Pancreatology,* 13**,** 48-57.

TSANG, A. C., PIRSHAHID, S. A., VIRGILI, G., GOTTLIEB, C. C., HAMILTON, J. & COUPLAND, S. G. 2015. Hydroxychloroquine and Chloroquine Retinopathy: A Systematic Review Evaluating the Multifocal Electroretinogram as a Screening Test. *Ophthalmology,* 122**,** 1239-1251. e4.

TSENG, D. S., VAN SANTVOORT, H. C., OFFERHAUS, G. J. A., BESSELINK, M. G., BOREL RINKES, I. H., VAN LEEUWEN, M. S. & MOLENAAR, I. Q. 2015. The role of CT in assessment of extra-regional lymph node involvement in pancreatic and peri-ampullary cancer: A prospective diagnostic accuracy study. *Pancreatology,* 15**,** S85-S86.

UIJL, S. G., LEIJTEN, F. S., PARRA, J., ARENDS, J. B., VAN HUFFELEN, A. C. & MOONS, K. G. 2005. What is the current evidence on decision-making after referral for temporal lobe epilepsy surgery?: A review of the literature. *Seizure,* 14**,** 534-540.

VAN DEN BRUEL, A., AERTGEERTS, B. & BUNTINX, F. 2006. Results of diagnostic accuracy studies are not always validated. *Journal of Clinical Epidemiology,* 59**,** 559.e1-559.e9.

VAN ERKEL, A. R. & PETER, M. 1998. Receiver operating characteristic (ROC) analysis: basic principles and applications in radiology. *European Journal of radiology,* 27**,** 88-94.

VAN TRIJFFEL, E., ANDEREGG, Q., BOSSUYT, P. & LUCAS, C. 2005. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual therapy,* 10**,** 256-269.

VILLAR, C. G. 2011. Evidence-based radiology for diagnostic imaging: What it is and how to practice it. *Radiología (English Edition),* 53**,** 326-334.

WALTHER, S., SCHUELER, S., TACKMANN, R., SCHUETZ, G. M., SCHLATTMANN, P. & DEWEY, M. 2014. Compliance with STARD checklist among studies of coronary CT angiography: systematic review. *Radiology,* 271**,** 74.

WANG, K., MUNIR, S., SHARIATI, K., ADAMS-WEBBER, T. & DORIA, A. 2015. Clinical utility of dual-energy x-ray absorptiometry for assessment of fractures in pediatric osteoporosis: Evidence-based knowledge synthesis. *Pediatric Radiology,* 45**,** S219-S221.

WANG, W., CHEN, L.-D., LU, M.-D., LIU, G.-J., SHEN, S.-L., XU, Z.-F., XIE, X.-Y., WANG, Y. & ZHOU, L.-Y. 2013. Contrast-enhanced ultrasound features of histologically proven focal nodular hyperplasia: diagnostic performance compared with contrast-enhanced CT. *European radiology,* 23**,** 2546-2554.

WARDLAW, J., CHAPPELL, F., BEST, J., WARTOLOWSKA, K. & BERRY, E. 2006. Non-invasive imaging compared with intra-arterial angiography in the diagnosis of symptomatic carotid stenosis: a meta-analysis. *The Lancet,* 367**,** 1503-1512.

WARDLAW, J. M., BRINDLE, W., CASADO, A. M., SHULER, K., HENDERSON, M., THOMAS, B., MACFARLANE, J., MANIEGA, S. M., LYMER, K. & MORRIS, Z. 2012. A systematic review of the utility of 1.5 versus 3 Tesla magnetic resonance brain imaging in clinical practice and research. *European radiology,* 22**,** 2295-2303.

WATERMAN, H., TILLEN, D., DICKSON, R. & DE KONING, K. 2000. Action research: a systematic review and guidance for assessment. *Health technology assessment (Winchester, England),* 5**,** iii-157.

WHITING, P., RUTJES, A. W., REITSMA, J. B., BOSSUYT, P. M. & KLEIJNEN, J. 2003. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol,* 3**,** 25.

WIDDIFIELD, J., LABRECQUE, J. & LIX, L. A Systematic Review to Evaluate the Quality and Reporting of Administrative Database Validation Studies for Rheumatic Diseases. JOURNAL OF RHEUMATOLOGY, 2011. J RHEUMATOL PUBL CO 365 BLOOR ST E, STE 901, TORONTO, ONTARIO M4W 3L4, CANADA, 1177-1178.

WIESKE, L., VERHAMME C, INHENFELDT, D., VAN DER SCHAAF, M., BOUWES, A. & SCHULTZ, M. 2012. Prediction of intensive care unit-aquired weakness using a simplified electrophysiological screening test. *Am J Respir Crit Care***,** 185.

WILCZYNSKI, N. L. 2008. Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication--before-and-after study. *Radiology,* 248**,** 817-823.

WILCZYNSKI, N. L. & HAYNES, R. B. 2007a. Indexing of Diagnostic Accuracy Studies in MEDLINE and EMBASE. *AMIA Annual Symposium Proceedings,* 2007**,** 801-805.

WILCZYNSKI, N. L. & HAYNES, R. B. Indexing of Diagnostic Accuracy Studies in MEDLINE and EMBASE. AMIA Annual Symposium Proceedings, 2007b. American Medical Informatics Association, 801.

WILCZYNSKI, N. L., MCKIBBON, K. A., WALTER, S. D., GARG, A. X. & HAYNES, R. B. 2013a. MEDLINE clinical queries are robust when searching in recent publishing years. *Journal of the American Medical Informatics Association : JAMIA,* 20**,** 363-368.

WILCZYNSKI, N. L., MCKIBBON, K. A., WALTER, S. D., GARG, A. X. & HAYNES, R. B. 2013b. MEDLINE clinical queries are robust when searching in recent publishing years. *Journal of the American Medical Informatics Association,* 20**,** 363-368.

WU, L., LI, Y., LI, Z., CAO, Y. & GAO, F. 2013. Diagnostic accuracy of narrow-band imaging for the differentiation of neoplastic from non-neoplastic colorectal polyps: a meta-analysis. *Colorectal Disease,* 15**,** 3-11.

YANG, D. H., KIM, Y.-H., ROH, J.-H., KANG, J.-W., HAN, D., JUNG, J., KIM, N., LEE, J. B., AHN, J.-M. & LEE, J.-Y. 2015. Stress Myocardial Perfusion CT in Patients Suspected of Having Coronary Artery Disease: Visual and Quantitative Analysis—Validation by Using Fractional Flow Reserve. *Radiology*, 141126.

ZAFAR, A., KHAN, G. I. & SIDDIQUI, M. 2008. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. *Clinical & experimental ophthalmology,* 36**,** 537-542.

ZHANG, L. J., WANG, Y., SCHOEPF, U. J., MEINEL, F. G., BAYER II, R. R., QI, L., CAO, J., ZHOU, C. S., ZHAO, Y. E. & LI, X. 2015. Image quality, radiation dose, and diagnostic accuracy of prospectively ECG-triggered high-pitch coronary CT angiography at 70 kVp in a clinical setting: comparison with invasive coronary angiography. *European radiology***,** 1-10.

Appendix 5: Literature Search 2003 – December 5, 2014

Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present> Search Strategy:

----

1 stard.ti,ab. (496)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (181)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (153)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (152)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (92)
6 or/1-5 (713)
7 (sensitiv$ or diagnos$).mp. (3204550)
8 di.fs. (2083205)
9 7 or 8 (4241507)
10 6 and 9 (502)
11 limit 10 to yr="2003 -Current" (463)
12 limit 11 to english language (440)
13 animal/ not human/ (4004891)
14 12 not 13 (434)

Database: EmbaseClassic+Embase<1947 to 2014 December 05> Search Strategy:

----

1 stard.ti,ab. (646)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (215)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (191)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (174)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (107)
6 or/1-5 (934)
7 di.fs. (2597849)
8 predict$.tw. (1229314)
9 specificity.tw. (409079)
10 diagnostic accuracy/ (186496)

11 (diagnostic adj1 accuracy).tw. (34059)
12 7 or 8 or 9 or 10 or 11 (3970285)
13 6 and 12 (482)
14 limit 13 to english language (448)
15 limit 14 to yr="2003 -Current" (427)
16 animal/ not human/ (1203732)
17 15 not 16 (423)

Date Run: 08/12/14 14:52:13.237

#1 stard:ti,ab,kw (Word variations have been searched) 60

#2 (reporting near/6 standard* near/6 diagnos*):ti,ab,kw (Word variations have been searched) 59

#3 (reporting near/6 standard* near/6 accuracy):ti,ab,kw (Word variations have been searched) 45

#4 (quality near/4 report* near/6 accuracy):ti,ab,kw (Word variations have been searched) 34

#5 (quality near/4 report* near/6 diagnos*):ti,ab,kw (Word variations have been searched) 38

#6 #1 or #2 or #3 or #4 or #5 96

Cochrane Methodology Register: Issue 3 of 4, July 2012**– 64 results**

Cochrane Central Register of Controlled Trials: Issue 11 of 12, November 2014 **– 23 results**

Cochrane Database of Systematic Reviews: Issue 12 of 12, December 2014 **– 5 results**

Database of Abstracts of Reviews of Effect: Issue 4 of 4, October 2014 **– 3 results**

Appendix 6: Literature Search 2003 – June 12, 2015

Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present> Search Strategy:

----

1 stard.ti,ab. (484)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (186)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (160)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (156)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (96)
6 or/1-5 (714)
7 (sensitiv$ or diagnos$).mp. (3216477)
8 di.fs. (2082892)
9 7 or 8 (4254220)
10 6 and 9 (505)
11 limit 10 to yr="2003 -Current" (465)
12 limit 11 to english language (444)
13 animal/ not human/ (3963496)

14 12 not 13 (434)

Database: EmbaseClassic+Embase<1947 to 2015 June 10> Search Strategy:

----

1 stard.ti,ab. (685)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (236)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (209)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (191)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (119)
6 or/1-5 (1003)
7 di.fs. (2678815)
8 predict$.tw. (1318150)
9 specificity.tw. (429590)
10 diagnostic accuracy/ (191130)
11 (diagnostic adj1 accuracy).tw. (36631)
12 7 or 8 or 9 or 10 or 11 (4143894)
13 6 and 12 (518)
14 limit 13 to english language (480)
15 limit 14 to yr="2003 -Current" (459)
16 animal/ not human/ (1254081)
17 15 not 16 (452)

Cochrane

#1 stard:ti,ab,kw (Word variations have been searched) 61

#2 (reporting near/6 standard* near/6 diagnos*):ti,ab,kw (Word variations have been searched) 61

#3 (reporting near/6 standard* near/6 accuracy):ti,ab,kw (Word variations have been searched) 46

#4 (quality near/4 report* near/6 accuracy):ti,ab,kw (Word variations have been searched) 34

#5 (quality near/4 report* near/6 diagnos*):ti,ab,kw (Word variations have been searched) 40

#6 #1 or #2 or #3 or #4 or #5 100

Cochrane Central Register of Controlled Trials : Issue 5 of 12, May 2015 = **24 results**

Cochrane Database of Systematic Reviews : Issue 6 of 12, June 2015 = **8 results**

Database of Abstracts of Reviews of Effect : Issue 2 of 4, April 2015 = **3 results**

Cochrane Methodology Register : Issue 3 of 4, July 2012 = **64 results**

Appendix 7: Revised Literature Search Strategy recommended by (Astin et al., 2008)

Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present> Search Strategy:

----

1 stard.ti,ab. (498)

2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (182)

3 (quality adj4 report$ adj6 diagnos$).ti,ab. (153)

4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (153)

5 (quality adj4 report$ adj6 accuracy).ti,ab. (94)

6 or/1-5 (718)

7 (sensitiv$ or diagnos$).mp. (3225034)

8 di.fs. (2096641)

9 7 or 8 (4268519)

10 6 and 9 (504)

11 limit 10 to yr="2003 -Current" (464)

12 limit 11 to english language (442)

13 false positive reactions/ or false negative reactions/ (35256)

14 (predictive adj4 value$).tw. (75763)

15 accura$.tw. (524738)

16 distinguish$.tw. (201673)

17 differentiat$.tw. (545624)

18 enhancement.tw. (166663)

19 detect$.tw. (1731268)

20 exp "Sensitivity and Specificity"/ (448746)

21 or/13-20 (3138050)

22 6 and 21 (337)

23 limit 22 to yr="2003 -Current" (320)

24 limit 23 to english language (301)

25 24 not 12 (49)

26 remove duplicates from 25 (42)

27 limit 6 to (english language and yr="2003 -Current") (649)

Additional bolded items inserted as recommended by (Astin et al., 2008)

Appendix 8: Literature Search 2003 – October 20, 2015 – third review of literature

Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present> Search Strategy:

--

1 stard.ti,ab. (505)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (200)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (169)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (168)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (99)
6 or/1-5 (752)
7 (sensitiv$ or diagnos$).mp. (3311768)

8 di.fs. (2148286)
9 7 or 8 (4383763)
10 6 and 9 (537)
11 limit 10 to yr="2003 -Current" (496)
12 limit 11 to english language (465)

13 2015$.ed,dc. (1371409)
14 201412$.ed,dc. (178570)
15 13 or 14 (1468720)
16 12 and 15 (53)

Database: Embase <1974 to 2015 October 20> Search Strategy:
----
1 stard.ti,ab. (713)
2 (reporting adj3 standard$ adj3 diagnos$).ti,ab. (250)
3 (quality adj4 report$ adj6 diagnos$).ti,ab. (225)
4 (reporting adj3 standard$ adj3 accuracy).ti,ab. (204)
5 (quality adj4 report$ adj6 accuracy).ti,ab. (124)
6 or/1-5 (1056)
7 di.fs. (2735485)
8 predict$.tw. (1365746)
9 specificity.tw. (431389)
10 diagnostic accuracy/ (194911)
11 (diagnostic adj1 accuracy).tw. (38002)
12 7 or 8 or 9 or 10 or 11 (4237298)
13 6 and 12 (551)
14 limit 13 to english language (512)
15 limit 14 to yr="2003 -Current" (491)
16 animal/ not human/ (1277760)
17 15 not 16 (484)

18 2015$.dd,em. (1777356)
19 201412$.dd,em. (129234)
20 18 or 19 (1868986)
21 17 and 20 (65) –

Cochrane

ID	Search	Hits
#1	stard:ti,ab,kw (Word variations have been searched)	61
#2	(reporting near/6 standard* near/6 diagnos*):ti,ab,kw (Word variations have been searched)	62
#3	(reporting near/6 standard* near/6 accuracy):ti,ab,kw (Word variations have been searched)	47
#4	(quality near/4 report* near/6 accuracy):ti,ab,kw (Word variations have been searched)	35
#5	(quality near/4 report* near/6 diagnos*):ti,ab,kw (Word variations have been searched)	41
#6	#1 or #2 or #3 or #4 or #5	**101 –**

Cochrane Database of Systematic Reviews : Issue 10 of 12, October 2015 = 8 results

Cochrane Central Register of Controlled Trials : Issue 9 of 12, September 2015 = 25 results

Database of Abstracts of Reviews of Effect : Issue 2 of 4, April 2015 = 3 results

Cochrane Methodology Register : Issue 3 of 4, July 2012 = 64 results

Appendix 9: Email Invitation to Radiologists

RE: An opportunity to participate in a research study titled “Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials”

Dear Doctor,
As part of my doctoral thesis, I am conducting a new study whereby you are being asked to participate in a needs assessment and potential development of a diagnostic accuracy tool that is specific to radiology. The purpose of the initial needs assessment is to come to a consensus on which items you deem are essential when interpreting the quality of diagnostic accuracy trials specific to radiology. Your level of expertise is a cornerstone in the success of this project. As an expert in the field your thoughts will help to fill in a gap that continues to exist. We estimate that 58 participants will be enrolled in the study from the Civic Campus of The Ottawa Hospital. The entire study will last approximately 2 years. Your participation in the study will last approximately 9-12 months.

The purpose of this study is to develop a tool that will benefit radiologists when interpreting diagnostic accuracy studies. This tool could potentially influence their practice going forward.

If you agree to participate in this research study your involvement will consist of participating in a needs assessment based on the current STARD (Standards for Reporting of Diagnostic Accuracy) tool. Your input will help to determine if further refinement to the tool is required to adapt it specifically for use in radiology. An initial draft version of the STARD-DI (Standards for Reporting of Diagnostic Accuracy Tool -Diagnostic Imaging) will be sent to you with a list of questions to determine which items you believe need to be included in the amended STARD – DI checklist. Each round of questions will be provided to you via the Delphi technique as this will allow for anonymity. This process will occur over 4 cycles. Each cycle will take about a month to complete, until a final version of the revised tool is created (if needed). The collected data from each set of questionnaires will be pooled and analyzed to develop the next set of questions for your input. Once a consensus has been met with the working group on the development of the revised tool, it will be piloted to the Residents and Fellows within the Department of Medical Imaging.

You may find the questions related to the needs assessment and Delphi technique redundant. You might not like all of the questions that you are asked. You do not have to answer any questions that make you uncomfortable. Although you may not receive any direct benefit from participating in this study, your involvement will greatly assist me in developing the requirements necessary to build this revised tool .This may benefit future trainees and radiologists in having a structured approach to interpreting the quality of diagnostic accuracy studies specific to radiology.

Your participation in the study is voluntary. If you decide to participate you have the right to withdraw consent at any point during the study without affecting your current or future employment at The Ottawa Hospital. If you withdraw your consent, the study team will no longer collect your personal identifying information for research purposes.

All information collected during your participation in this study will be de-identified and will not contain information that identifies you, such as your name or email address. The link between your needs assessment responses and your name and contact information will be stored securely and separate from your study records, and will not leave The Ottawa Hospital. All paper records will be stored securely in a locked file and/or office. All electronic records will be protected by a user password which is only accessible to myself. Any documents leaving The Ottawa Hospital will contain only anonymous data. This includes publications or presentations resulting from this study. All personal information will be kept confidential, unless release is required by law. Representatives of the Ottawa Health Science Network Research Ethics Board (OHSN-REB) as well as the Ottawa Hospital Research Institute may review your original study records, for audit purposes only, under my supervision. Research records will be kept for 10 years, after this time they will be destroyed.

If you have any questions about the study, please do not hesitate to contact me by responding to this email (baschwarz@toh.on.ca) or at 613-798-5555 ext. 17522.

The Ottawa Health Science Network Research Ethics Board (OHSN-REB) has reviewed the plans for this research study. The Board considers the ethical aspects of all research studies involving human participants at The Ottawa Hospital. If you have any questions about your rights as a study participant, you may contact the Chairperson at 613-798-5555, extension 16719.

Please note that there will be no written consent for this study. Providing your responses to the needs assessment based on the current STARD diagnostic accuracy tool implies your consent to participate in this research study. Please indicate your interest to participate by responding to this email sent to you in confidence.
Yours Truly,

Betty Anne Schwarz
Principal Investigator
Co-Investigator: Dr. Wael Shabana
Middlesex University Supervisor: Dr. Barbara Workman

Appendix 10: STARD_DI Checklist

STARD-DI checklist for the reporting of radiology diagnostic accuracy trials			
Section & Topic	Item #		Page Number
Abstract, Title	1	Provide the name or a phrase that describes the intervention.	
Introduction	2	Identify the aim, describe the theory (hypothesis), or goal (diagnostic accuracy) essential to the intervention.	

STARD-DI checklist for the reporting of radiology diagnostic accuracy trials			
Section & Topic	Item #		Page Number
METHODS			
Participants	3	Was the inclusion/exclusion/ disease severity/ diagnostic criteria provided for the study population?	
	4	Was recruitment based on disease entity?	
	(i)	If not a pilot study: was there a power analysis provided?	
	(ii)	Was the number of study population denominator and the subgroup? – i.e., how many were benign or malignant.	
	(iii)	Was the population number provided of those that received the index test and reference standard?	
	5	Were the participants enrolled prospectively (randomized) or was it a retrospective analysis?	
	6	Was the location provided where the testing occurred as well as any necessary infrastructure?	
DATA COLLECTION - TECHNIQUE			
Test Methods	7	Describe the reference standard and it's rationale.	
	(i)	Was the technical background provided? Is it sound?	
	(ii)	Was the nature of the disease entity or lesion that needed to diagnose adequately described?	
	(iii)	If technique is not standard of care, was imaging protocol provided?	
	8	Were the technical specifications, rationale for use and cut-off values to diagnose provided for the index test and the reference standard?	
	(i)	If the test was modified during the study were the changes provided?	
	9	Was the training level of the investigator provided? Was there any extra training needed? i.e. new technique	
	(ii)	How were the discrepancies in the data dealt with?	
Statistical Methods	10	Was the statistical test applied and used correctly?	
RESULTS			

STARD-DI checklist for the reporting of radiology diagnostic accuracy trials			
Section & Topic	Item #		Page Number
Participants	11	Were patient demographics provided including age, sex, presenting diagnosis or symptoms? Were any co-morbidities, and concurrent therapies provided?	
	(i)	Was a flow diagram provided to illustrate allotment of testing?	
Test Results	12	Was the time provided for the reference test and index test provided as well as any other treatment provided in between the two tests?	
	13	Was the severity of the disease entity described and or diagnostic criterion provided?	
	15	Were adverse events reported for either the index test or reference standard?	
Estimates	16	Were the estimates of diagnostic accuracy (kappa) and measures of statistical uncertainty provided (95% CI)?	
	17	How were indeterminate and missing results handled for the index test and any outliers?	
	18	Were estimates of interobserver variability for diagnostic accuracy between the readers reported? Were multiple sites reported (RCT)?	
	19	Were estimates of test reproducibility reported (kappa)?	
DISCUSSION	20	Was the clinical significance of the study findings provided?	

STARD Tool, 2003, TIDIER Checklist not yet published, 2013.

Appendix 11: Email invitation to send to radiological experts who agree to participate in my study

RE: An opportunity to participate in a research study titled “Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials”
Dear Doctor,

Please note that this email has been sent to you in confidence.

Thank you for agreeing to participate in this research study which aims to develop a tool that will benefit radiologists when interpreting diagnostic accuracy studies.

Please see the attached draft tool STARD-DI (Standards for Reporting of Diagnostic Accuracy - Diagnostic Imaging). Based on your expertise and perspective can you please advice which items you think should be included in the development of a revised tool for clinicians to use as a guide when interpreting radiology diagnostic accuracy trials?

Your input will help to determine if further refinement to the tool is required to adapt it specifically for use in radiology. This process will occur over 4 cycles until a final version of the revised tool is created (if needed).

Your participation in the study is voluntary. If you decide to participate you have the right to withdraw consent at any point during the study without affecting your current or future employment at The Ottawa Hospital. If you withdraw your consent, the study team will no longer collect your personal identifying information for research purposes. Please note that there will be no written consent for this study. Providing your responses to the needs assessment based on the current STARD-DI diagnostic accuracy tool implies your consent to participate in this research study. The entire study will last approximately two years. Your participation in the study will last approximately 9-12 months. Should you have any questions, please do not hesitate to contact me.

Thanks again,

Betty Anne

Betty Anne Schwarz MSc, BA, RN, Doctoral Candidate
Medical Imaging Research Program Manager
502- 751 Parkdale Ave
Ottawa, ON K1Y 1J7
(613) 798-5555 ext. 17522
baschwarz@toh.on.ca

Appendix 12: STARD checklist for reporting of studies of diagnostic accuracy

(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS			

Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	
	22	How indeterminate results, missing data and outliers of the index tests were handled.	

	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

Appendix 13: Permission to Present to Radiology Residents

Would you need written consent from the residents for their participation?

Schedule-wise, we have an opening on June 23 at 4pm. Matt, this would fall right after your research session with the residents.

Rebecca

On Jun 11, 2015, at 2:08 PM, Schwarz, Betty Anne <baschwarz@toh.on.ca> wrote:

Dear Rebecca,

I am contacting you to ask for your permission to speak to the residents (later on this summer) for 10-15 minutes at their academic half-day about my research.
The following is my Abstract that was submitted to RSNA which provides a summary of where I'm at.

Scientific Presentation

Project Title: Developing a Standardized Tool for the Interpretation and Reporting of Radiology Diagnostic Accuracy Studies (RADART)

Purpose: Radiologists rely on the results of published clinical trials within the literature as a guide to achieving standards of accuracy. However; this is predicated on the fact that the quality of reporting for diagnostic accuracy trials is of high standards, which it is not. The field of radiology needs to develop standardised methods of interpreting research that is accurate, transferrable and appropriate to patient care.

Materials and Methods: Although the STARD (Standards for the Reporting of Diagnostic Accuracy Studies) tool was developed in 2003 to improve the quality of reporting diagnostic accuracy studies, its use has been slowly adopted whereby revisions may be necessary. The goal of this project is to develop a standardized tool that is specific to radiology diagnostic accuracy trials for radiologists and trainees to use when reporting and interpreting radiology diagnostic accuracy studies. This action research project is currently being developed and tested.

Results: The revised tool (RADART) has been developed with radiological experts via the Delphi technique. The explanation and elaboration document is being developed which will accompany the RADART as a user guideline. It will be piloted with the radiology residents and Fellows in comparison to the current STARD. This is the first time a reporting tool will be validated in practice.

Conclusions: By developing a revised tool and reporting guideline that is specific to interpreting the quality of reporting of radiology diagnostic accuracy studies it will enable the clinician to interpret and evaluate the reliability and usefulness of the results specifically to radiology.

Characters with Spaces: 1,802

In order to test the RADART may I please present the RADART to the residents at Academic half day?

This presentation would take 10-15 minutes – it would include a brief overview of the STARD and how the RADART was developed (Delphi technique with radiologists and was REB approved)

I would define what was different about the RADART (how it is specific to Radiology) and explain that an article would be sent to them for review using the RADART.

I have developed an Elaboration document which would be provided to them as well. I would also send them the STARD tool and elaboration document as a reference only in comparison to the RADART. I also wrote a 2 page summary for each item of the RADART that summarizes the item and indicates which item # from the STARD I am referring to. Some items are new to the RADART and some items from the STARD were not included as they did not meet consensus by the cohort.

Once they had read the article and familiarized themselves with RADART they would be asked to complete a survey via Survey Monkey indicating their level of confidence interpreting the article based on the item provided in the RADART checklist. This would consist of a Likert Scale 1-10 (1 least confident – 10 most confident). The RADART is 25 items long – same as STARD.

To eliminate any biases the RADART and elaboration document would only be sent to the residents and Fellows.

Then I would like to meet with them once more in follow-up to provide them the responses from Survey Monkey and to ask them how useful they found the RADART – this would consist of short questionnaire (paper and pen) with space for the residents to insert their comments. I would leave a box in the room for them to leave their response sheet upon leaving the room.

Please contact me with any questions or concerns you may have.

I am very excited to be at this stage of my Doctoral study and I welcome sharing my results with the department once data collection is completed.

Thanks very much,

Betty Anne

Appendix 14: Permission to Present to the Radiology Fellows

Certainly yes Betty Anne

I will copy Avril and she can help you select a good date together

Best John

Dear John,

I am contacting you to ask for your permission to speak to the Fellows (later on this summer) for 10-15 minutes at their academic half-day about my research.

The following is my Abstract that was submitted to RSNA which provides a summary of where I'm at.

Scientific Presentation
Project Title: Developing a Standardized Tool for the Interpretation and Reporting of Radiology Diagnostic Accuracy Studies (RADART)
Purpose: Radiologists rely on the results of published clinical trials within the literature as a guide to achieving standards of accuracy. However; this is predicated on the fact that the quality of reporting for diagnostic accuracy trials is of high standards, which it is not. The field of radiology needs to develop standardised methods of interpreting research that is accurate, transferrable and appropriate to patient care.

Materials and Methods: Although the STARD (Standards for the Reporting of Diagnostic Accuracy Studies) tool was developed in 2003 to improve the quality of reporting diagnostic accuracy studies, its use has been slowly adopted whereby revisions may be necessary. The goal of this project is to develop a standardized tool that is specific to radiology diagnostic accuracy trials for radiologists and trainees to use when reporting and interpreting radiology diagnostic accuracy studies. This action research project is currently being developed and tested.

Results: The revised tool (RADART) has been developed with radiological experts via the Delphi technique. The explanation and elaboration document is being developed which will accompany the RADART as a user guideline. It will be piloted with the radiology residents and Fellows in comparison to the current STARD. This is the first time a reporting tool will be validated in practice.

Conclusions: By developing a revised tool and reporting guideline that is specific to interpreting the quality of reporting of radiology diagnostic accuracy studies it will enable the clinician to interpret and evaluate the reliability and usefulness of the results specifically to radiology.

Characters with Spaces: 1,802

In order to test the RADART may I please present the RADART to the Fellows at their Academic half day?
This presentation would take 10-15 minutes – it would include a brief overview of the STARD and how the RADART was developed (Delphi technique with radiologists and was REB approved).

I would define what was different about the RADART (how it is specific to Radiology) and explain that an article would be sent to them for review using the RADART.

I have developed an Elaboration document which would be provided to them as well. I would also send them the STARD tool and elaboration document as a reference only in comparison to the RADART. I also wrote a 2 page summary for each item of the RADART that summarizes the item and indicates which item # from the STARD I am referring to. Some items are new to the RADART and some items from the STARD were not included as they did not meet consensus by the cohort.

Once they had read the article and familiarized themselves with RADART they would be asked to complete a survey via Survey Monkey indicating their level of confidence interpreting the article based on the item provided in the RADART checklist. This would consist of a Likert Scale 1-10 (1 least confident – 10 most confident). The RADART is 25 items long – same as STARD.

To eliminate any biases the RADART and elaboration document would only be sent to the residents and Fellows. Then I would like to meet with them once more in follow-up to provide them the responses from Survey Monkey and to ask them how useful they found the RADART – this would consist of short questionnaire (paper and pen) with space for the Fellows to insert their comments. I would leave a box in the room for them to leave their response sheet upon leaving the room.

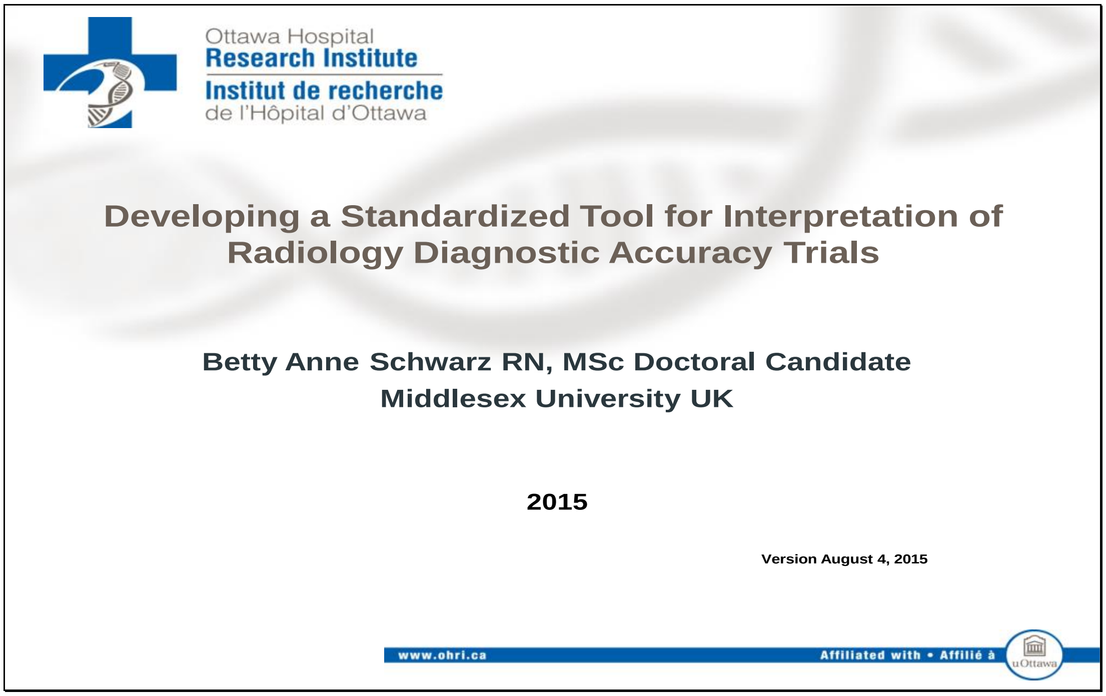
I realize that the Fellows don't all necessarily meet at the same time so I am more than willing to present to them on multiple occasions pending their academic schedule.

Please contact me with any questions or concerns you may have.
I am very excited to be at this stage of my Doctoral study and I welcome sharing my results with the department once data collection is completed.
Thanks very much,

Betty Anne

Appendix 15: Power-Point Presentation Introduction to Study

Slide 1



Ottawa Hospital
Research Institute
Institut de recherche
de l'Hôpital d'Ottawa

Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials

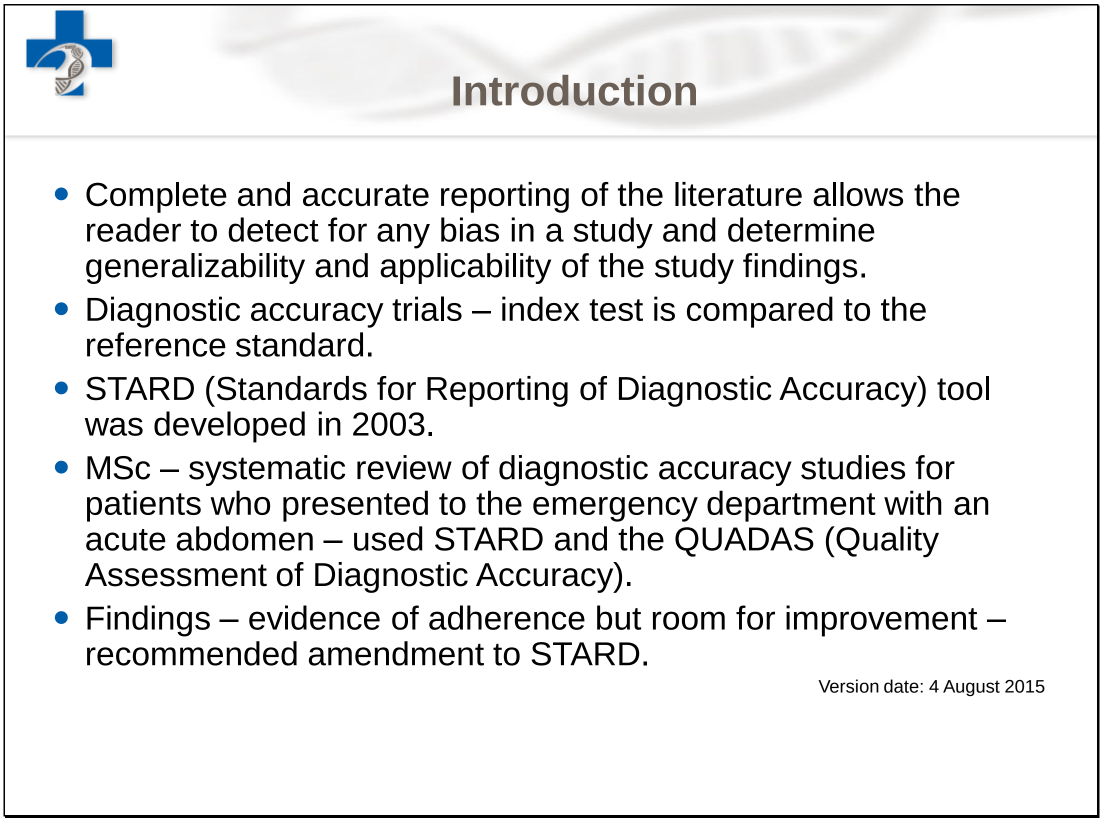
Betty Anne Schwarz RN, MSc Doctoral Candidate
Middlesex University UK

2015

Version August 4, 2015

www.ohri.ca Affiliated with • Affilié à uOttawa

Slide 2

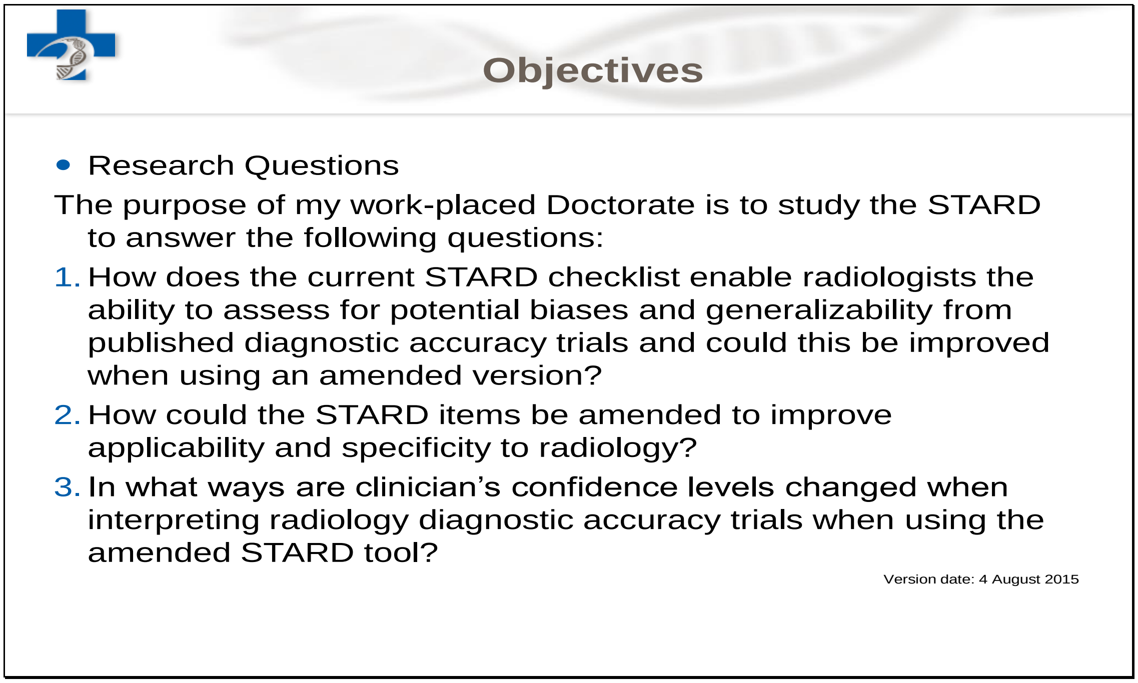


Introduction

- Complete and accurate reporting of the literature allows the reader to detect for any bias in a study and determine generalizability and applicability of the study findings.
- Diagnostic accuracy trials – index test is compared to the reference standard.
- STARD (Standards for Reporting of Diagnostic Accuracy) tool was developed in 2003.
- MSc – systematic review of diagnostic accuracy studies for patients who presented to the emergency department with an acute abdomen – used STARD and the QUADAS (Quality Assessment of Diagnostic Accuracy).
- Findings – evidence of adherence but room for improvement – recommended amendment to STARD.

Version date: 4 August 2015

Slide 3



Objectives

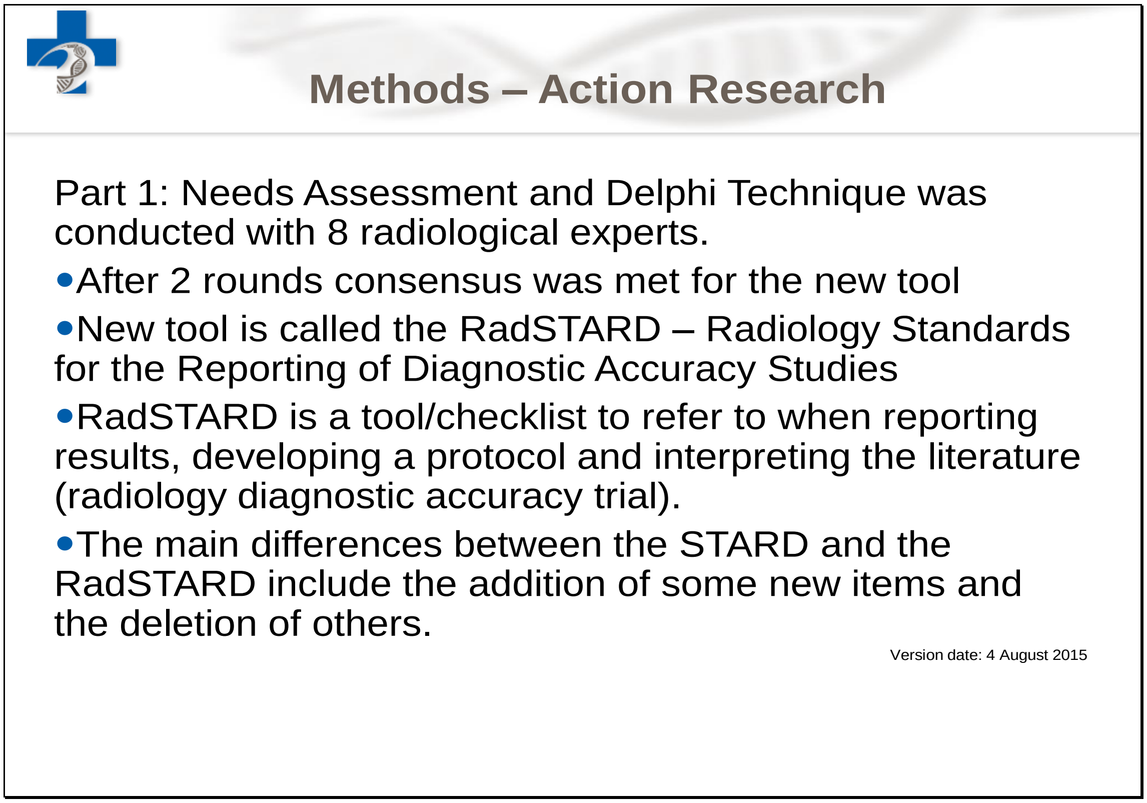
- Research Questions

The purpose of my work-placed Doctorate is to study the STARD to answer the following questions:

1. How does the current STARD checklist enable radiologists the ability to assess for potential biases and generalizability from published diagnostic accuracy trials and could this be improved when using an amended version?
2. How could the STARD items be amended to improve applicability and specificity to radiology?
3. In what ways are clinician's confidence levels changed when interpreting radiology diagnostic accuracy trials when using the amended STARD tool?

Version date: 4 August 2015

Slide 4



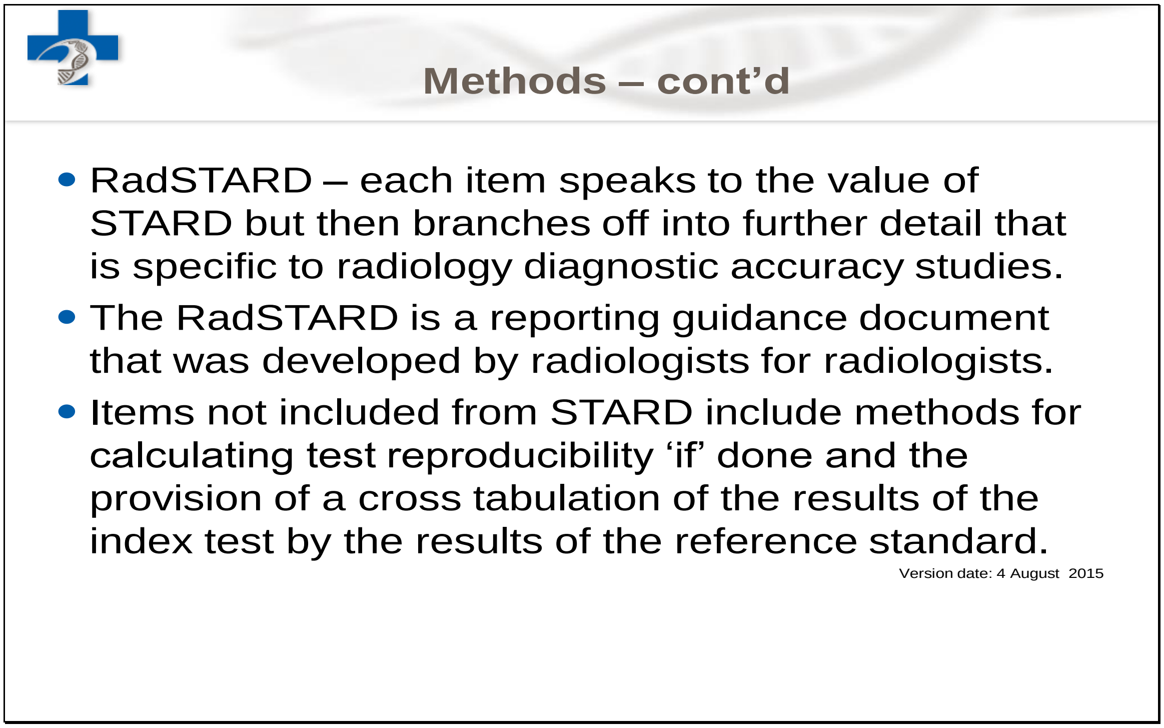
Methods – Action Research

Part 1: Needs Assessment and Delphi Technique was conducted with 8 radiological experts.

- After 2 rounds consensus was met for the new tool
- New tool is called the RadSTARD – Radiology Standards for the Reporting of Diagnostic Accuracy Studies
- RadSTARD is a tool/checklist to refer to when reporting results, developing a protocol and interpreting the literature (radiology diagnostic accuracy trial).
- The main differences between the STARD and the RadSTARD include the addition of some new items and the deletion of others.

Version date: 4 August 2015

Slide 5

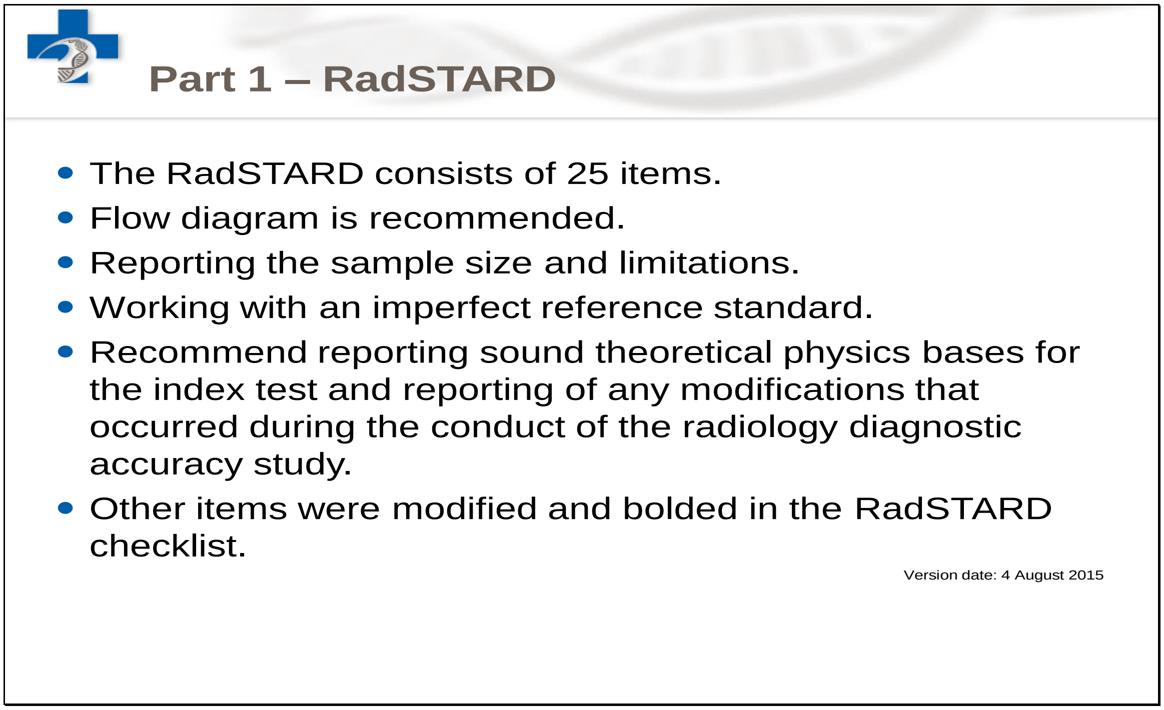


Methods – cont'd

- RadSTARD – each item speaks to the value of STARD but then branches off into further detail that is specific to radiology diagnostic accuracy studies.
- The RadSTARD is a reporting guidance document that was developed by radiologists for radiologists.
- Items not included from STARD include methods for calculating test reproducibility 'if' done and the provision of a cross tabulation of the results of the index test by the results of the reference standard.

Version date: 4 August 2015

Slide 6

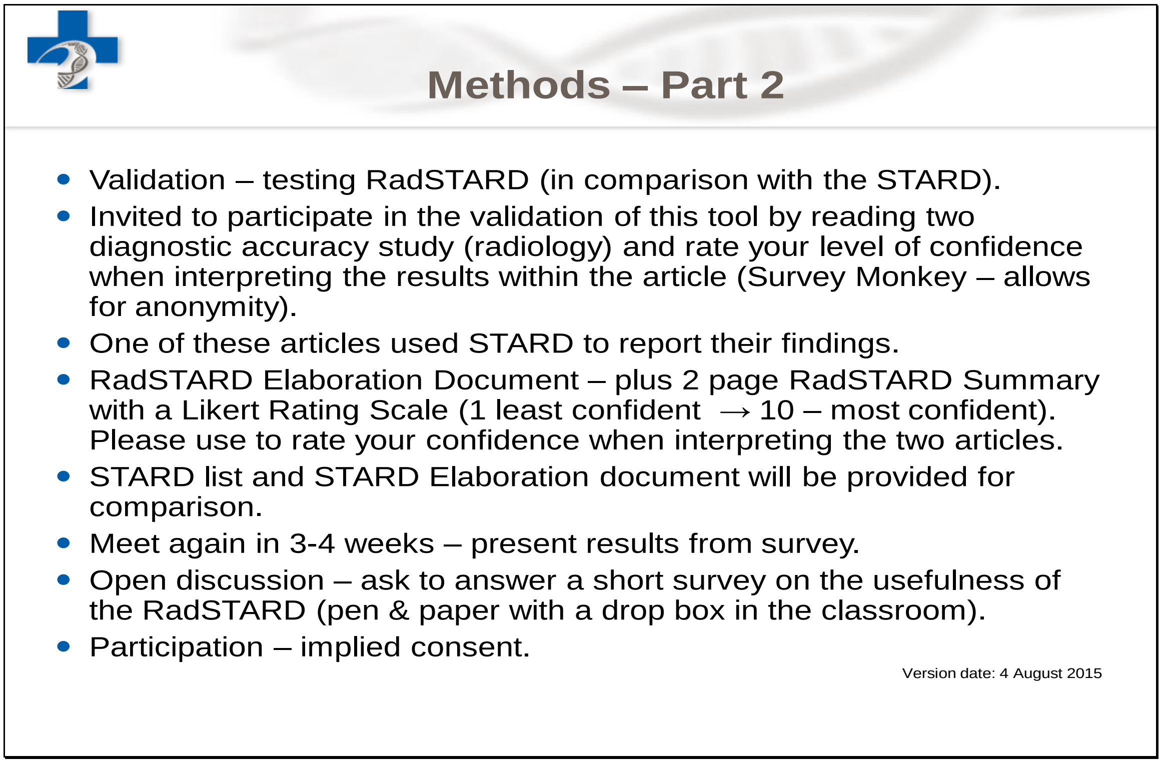


Part 1 – RadSTARD

- The RadSTARD consists of 25 items.
- Flow diagram is recommended.
- Reporting the sample size and limitations.
- Working with an imperfect reference standard.
- Recommend reporting sound theoretical physics bases for the index test and reporting of any modifications that occurred during the conduct of the radiology diagnostic accuracy study.
- Other items were modified and bolded in the RadSTARD checklist.

Version date: 4 August 2015

Slide 7

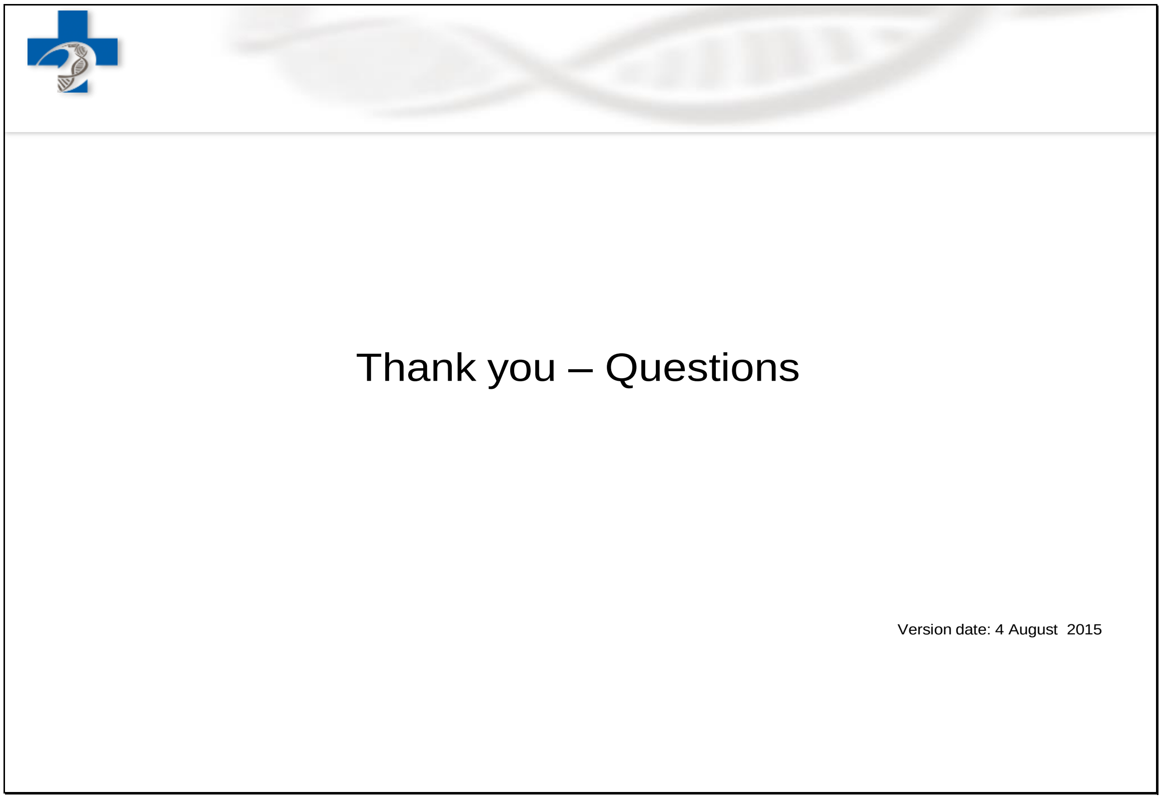


Methods – Part 2

- Validation – testing RadSTARD (in comparison with the STARD).
- Invited to participate in the validation of this tool by reading two diagnostic accuracy study (radiology) and rate your level of confidence when interpreting the results within the article (Survey Monkey – allows for anonymity).
- One of these articles used STARD to report their findings.
- RadSTARD Elaboration Document – plus 2 page RadSTARD Summary with a Likert Rating Scale (1 least confident → 10 – most confident). Please use to rate your confidence when interpreting the two articles.
- STARD list and STARD Elaboration document will be provided for comparison.
- Meet again in 3-4 weeks – present results from survey.
- Open discussion – ask to answer a short survey on the usefulness of the RadSTARD (pen & paper with a drop box in the classroom).
- Participation – implied consent.

Version date: 4 August 2015

Slide 8

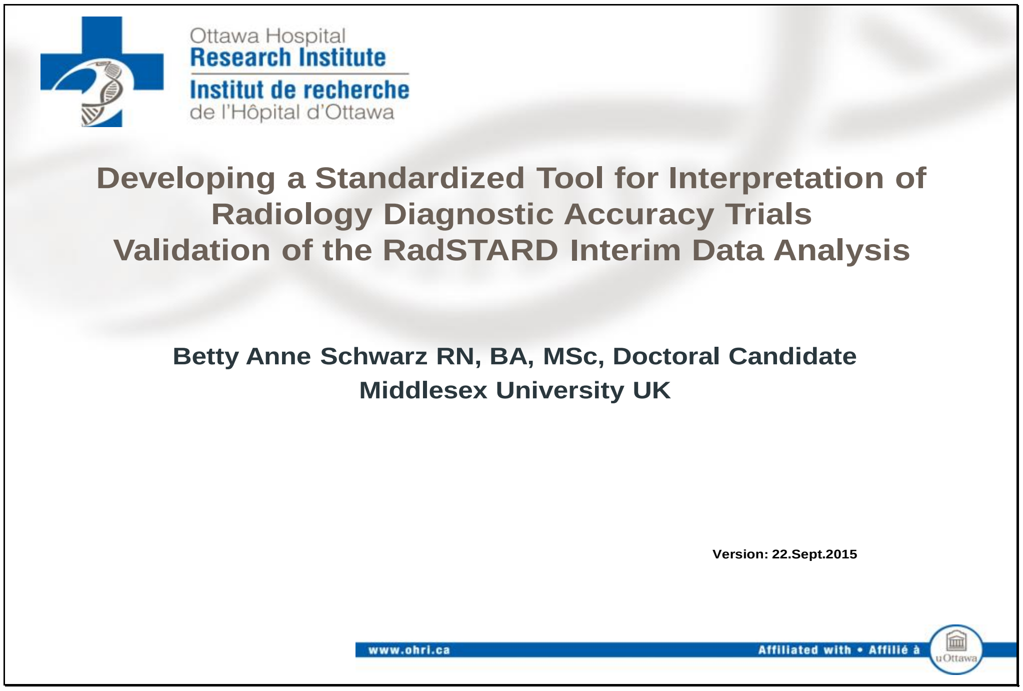


Thank you – Questions

Version date: 4 August 2015

Appendix 16: Power-Point Presentation – Interim Analyses

Slide 1



Ottawa Hospital
Research Institute
Institut de recherche
de l'Hôpital d'Ottawa

Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials
Validation of the RadSTARD Interim Data Analysis

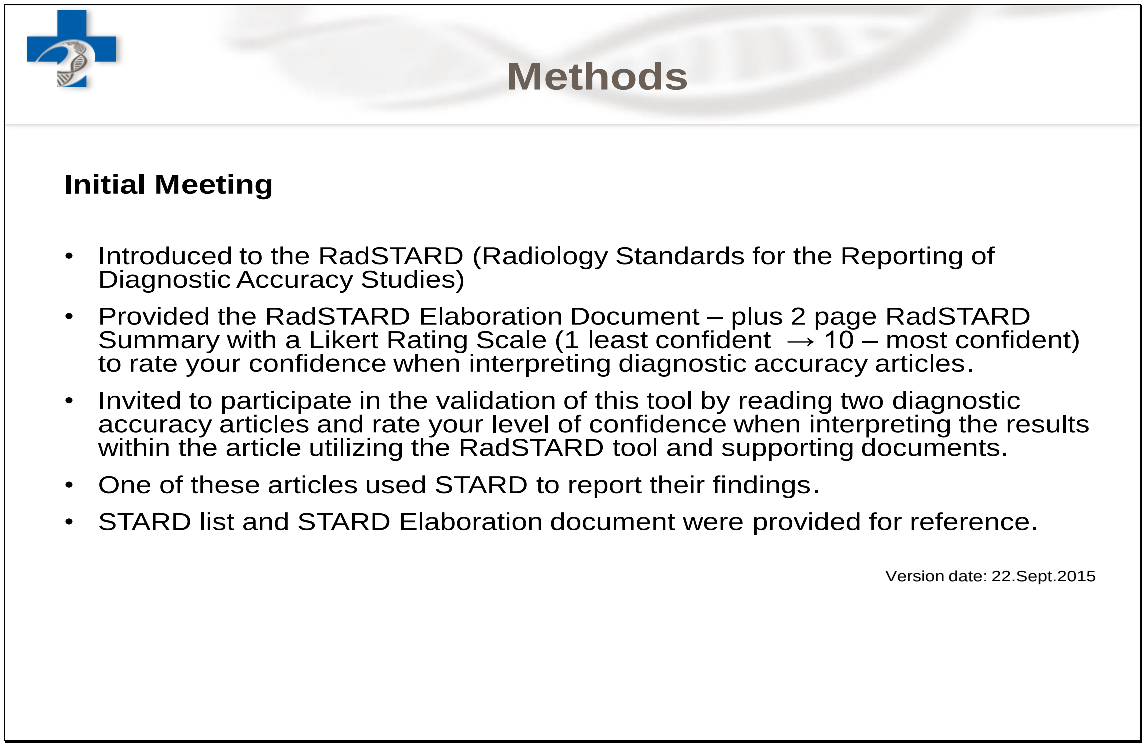
Betty Anne Schwarz RN, BA, MSc, Doctoral Candidate
Middlesex University UK

Version: 22.Sept.2015

www.ohri.ca

Affiliated with • Affilié à uOttawa

Slide 2



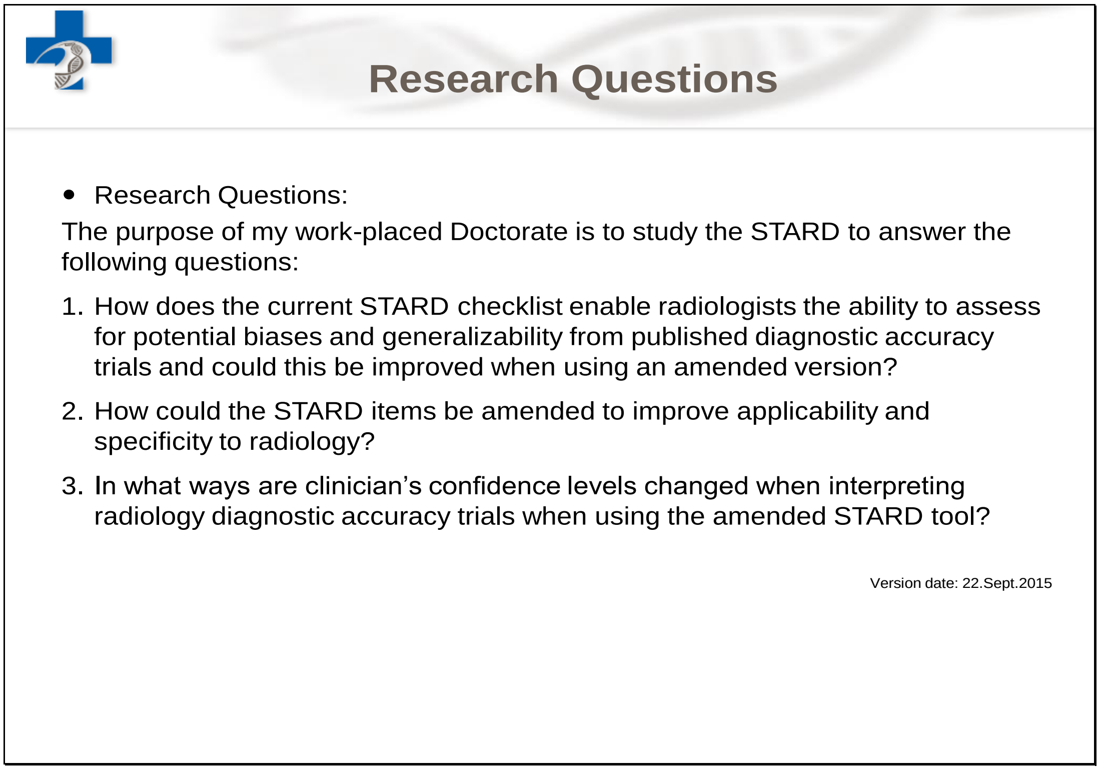
Methods

Initial Meeting

- Introduced to the RadSTARD (Radiology Standards for the Reporting of Diagnostic Accuracy Studies)
- Provided the RadSTARD Elaboration Document – plus 2 page RadSTARD Summary with a Likert Rating Scale (1 least confident → 10 – most confident) to rate your confidence when interpreting diagnostic accuracy articles.
- Invited to participate in the validation of this tool by reading two diagnostic accuracy articles and rate your level of confidence when interpreting the results within the article utilizing the RadSTARD tool and supporting documents.
- One of these articles used STARD to report their findings.
- STARD list and STARD Elaboration document were provided for reference.

Version date: 22.Sept.2015

Slide 3



Research Questions

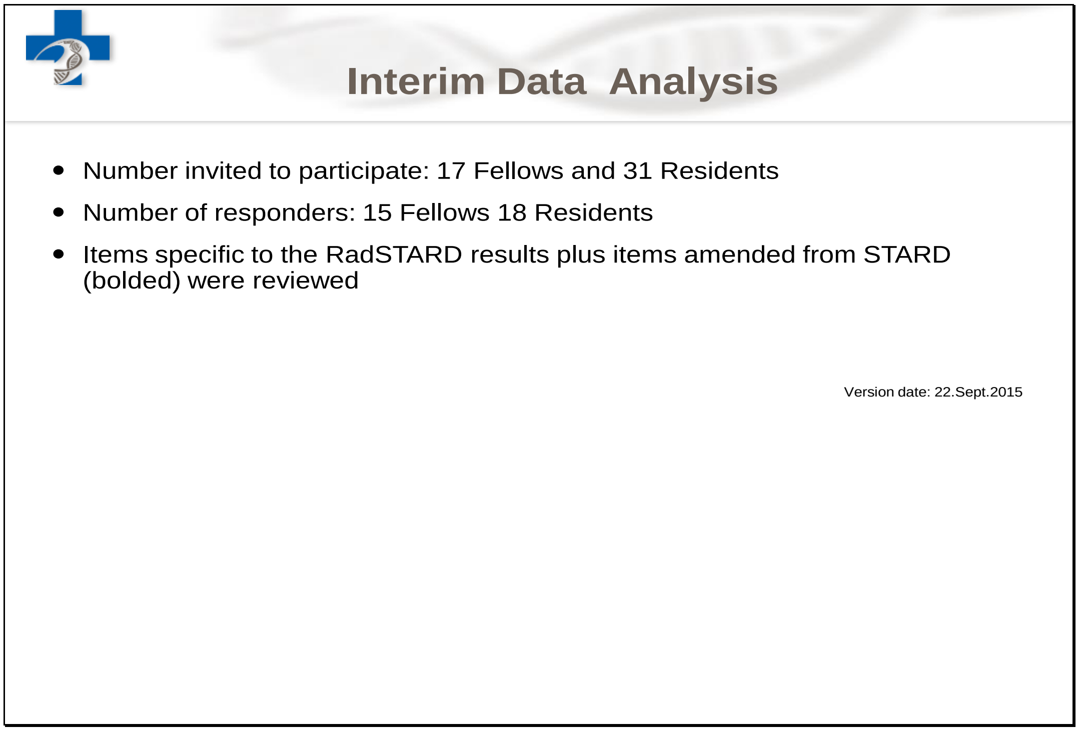
- Research Questions:

The purpose of my work-placed Doctorate is to study the STARD to answer the following questions:

1. How does the current STARD checklist enable radiologists the ability to assess for potential biases and generalizability from published diagnostic accuracy trials and could this be improved when using an amended version?
2. How could the STARD items be amended to improve applicability and specificity to radiology?
3. In what ways are clinician's confidence levels changed when interpreting radiology diagnostic accuracy trials when using the amended STARD tool?

Version date: 22.Sept.2015

Slide 4

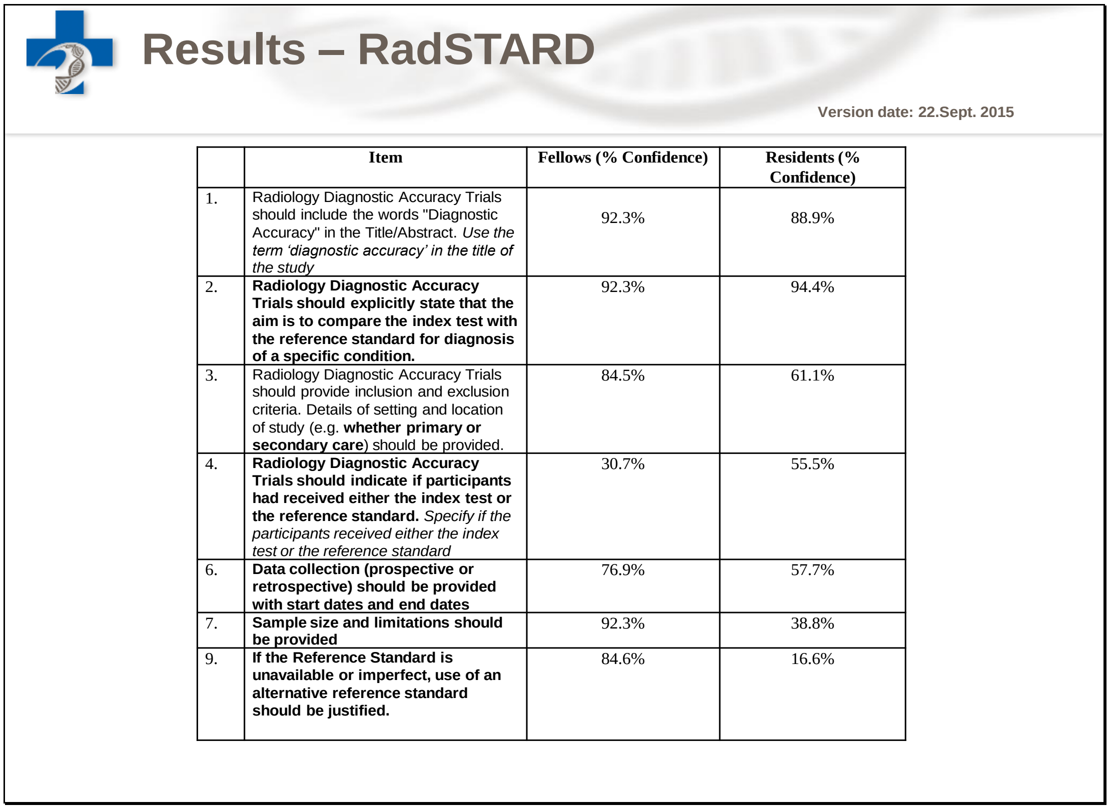


Interim Data Analysis

- Number invited to participate: 17 Fellows and 31 Residents
- Number of responders: 15 Fellows 18 Residents
- Items specific to the RadSTARD results plus items amended from STARD (bolded) were reviewed

Version date: 22.Sept.2015

Slide 5

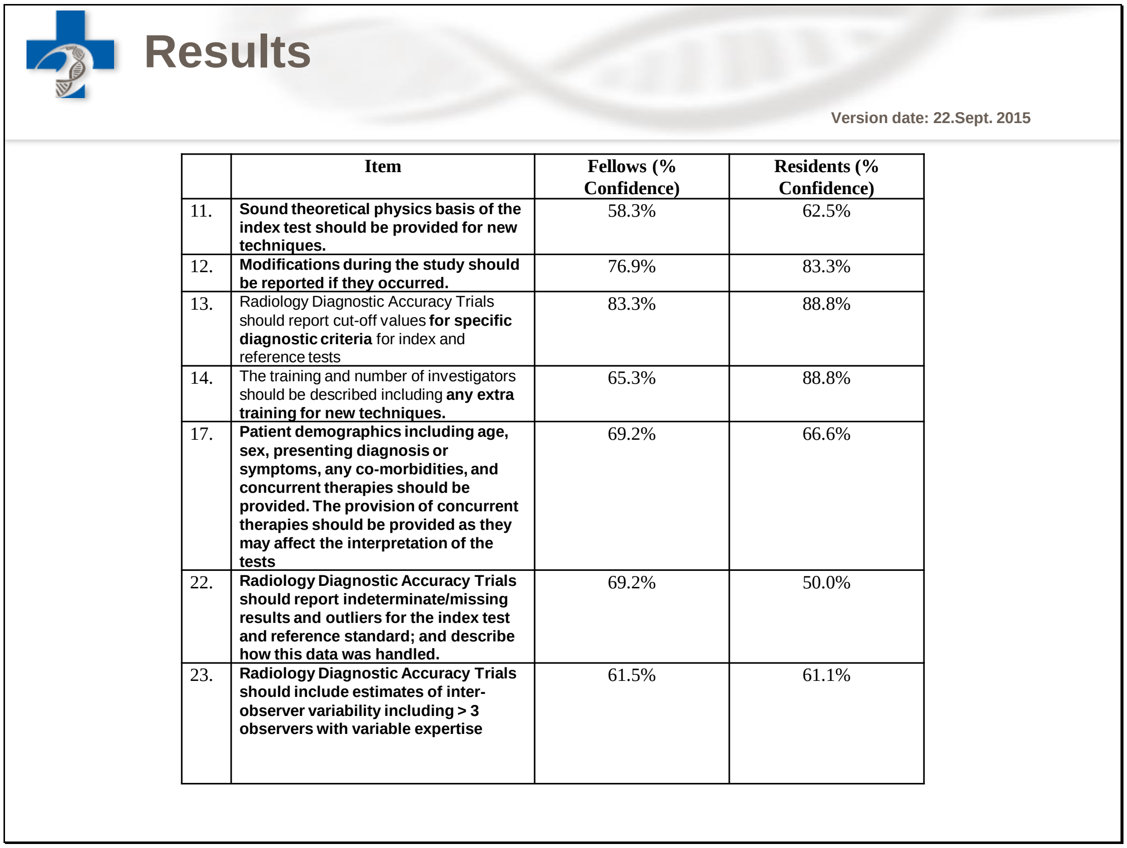


Results – RadSTARD

Version date: 22.Sept. 2015

	Item	Fellows (% Confidence)	Residents (% Confidence)
1.	Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract. *Use the term 'diagnostic accuracy' in the title of the study*	92.3%	88.9%
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare the index test with the reference standard for diagnosis of a specific condition.**	92.3%	94.4%
3.	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided.	84.5%	61.1%
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.** *Specify if the participants received either the index test or the reference standard*	30.7%	55.5%
6.	**Data collection (prospective or retrospective) should be provided with start dates and end dates**	76.9%	57.7%
7.	**Sample size and limitations should be provided**	92.3%	38.8%
9.	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be justified.**	84.6%	16.6%

Slide 6

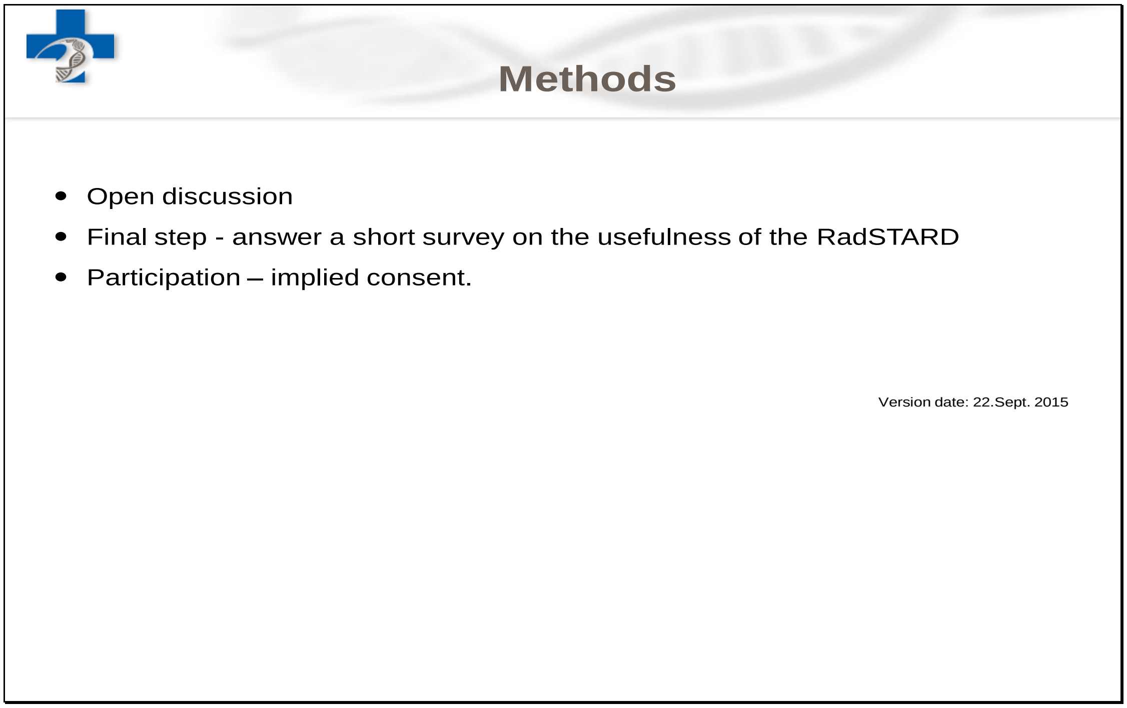


Results

Version date: 22.Sept. 2015

	Item	Fellows (% Confidence)	Residents (% Confidence)
11.	**Sound theoretical physics basis of the index test should be provided for new techniques.**	58.3%	62.5%
12.	**Modifications during the study should be reported if they occurred.**	76.9%	83.3%
13.	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests	83.3%	88.8%
14.	The training and number of investigators should be described including **any extra training for new techniques.**	65.3%	88.8%
17.	**Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies should be provided. The provision of concurrent therapies should be provided as they may affect the interpretation of the tests**	69.2%	66.6%
22.	**Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers for the index test and reference standard; and describe how this data was handled.**	69.2%	50.0%
23.	**Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability including > 3 observers with variable expertise**	61.5%	61.1%

Slide 7

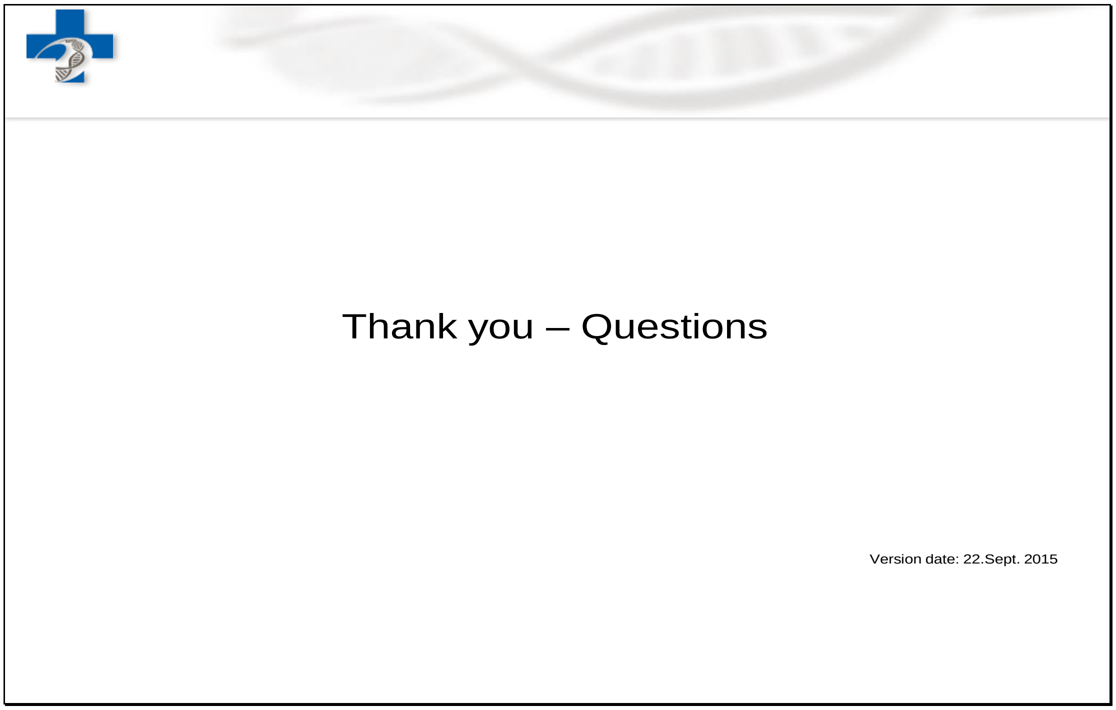


Methods

- Open discussion
- Final step - answer a short survey on the usefulness of the RadSTARD
- Participation – implied consent.

Version date: 22.Sept. 2015

Slide 8



Thank you – Questions

Version date: 22.Sept. 2015

Appendix 17: Email Invitation and Consent mailed to Radiology Residents and Fellows

Includes two articles: RadSTARD tool and elaboration document, RadSTARD 2 page summary, STARD tool, STARD elaboration document)

RE: An opportunity to participate in a research study titled “Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials”

Dear Doctor,

As part of my doctoral thesis, I am conducting a new study whereby you are being asked to participate in the assessment and usefulness of a newly developed reporting tool for radiology diagnostic accuracy studies. This reporting tool is called the RadSTARD (Radiology Standards for the Reporting of Diagnostic Accuracy Studies). We estimate that 58 participants will be enrolled in the study from the Civic Campus of The Ottawa Hospital. The entire study will last approximately 2 years. Your participation in the study will last approximately 2 months.

The purpose of this study is to develop a tool that will benefit radiologists when interpreting diagnostic accuracy studies. This tool could potentially influence their practice going forward.

If you agree to participate in this research study your involvement will consist of reading two radiology diagnostic accuracy articles whereby you will be asked to rate your confidence in interpreting the article as per each item of the RadSTARD tool through an online questionnaire on Survey Monkey. This questionnaire has 25 questions and should take you approximately 10 minutes to complete. You will be provided with the RadSTARD Elaboration Document plus a 2 page summary of the RadSTARD tool. In addition to these documents you will be provided with a copy of the current STARD tool and STARD Elaboration Document as a reference when comparing to the RadSTARD. The collected data from Survey Monkey will be pooled and analysed to present back to you at another Academic teaching session. This may occur in the next 3-4 week’s time at one of your next Academic half days. At that time you may be asked to complete another short questionnaire on the usefulness of the RadSTARD tool.

You might not like all of the questions that you are asked. You do not have to answer any questions that make you uncomfortable.

Although you may not receive any direct benefit from participating in this study, your involvement will greatly assist me in validating this revised tool. This may benefit future trainees and radiologists in having a structured approach to interpreting the quality of diagnostic accuracy studies specific to radiology.

Your participation in the study is voluntary. If you decide to participate you have the right to withdraw consent at any point during the study without affecting your current or future employment at The Ottawa Hospital. You can withdraw from the study by not submitting or completing the survey. Once the survey has been submitted I cannot withdraw your data from the study as participation anonymization will have already occurred.

All information collected during your participation in this study is anonymous and will not contain information that identifies you, such as your name or email address. All paper records will be stored securely in a locked file and/or office. All electronic records will be protected by a user password which is only accessible to me. Any documents leaving The Ottawa Hospital will contain only anonymous data. This includes publications or presentations resulting from this study. All personal information will be kept confidential, unless release is required by law. Representatives of the Ottawa Health Science Network Research Ethics Board (OHSN-REB) as well as the Ottawa Hospital Research Institute may review your original study records, for audit purposes only, under my supervision. Research records will be kept for 10 years, after this time they will be destroyed.

If you have any questions about the study, please do not hesitate to contact me by responding to this email (baschwarz@toh.on.ca) or at 613-798-5555 ext. 17522.

The Ottawa Health Science Network Research Ethics Board (OHSN-REB) has reviewed the plans for this research study. The Board considers the ethical aspects of all research studies involving human participants at The Ottawa Hospital. If you have any questions about your rights as a study participant, you may contact the Chairperson at 613-798-5555, extension 16719.

Please note that there will be no written consent for this study. Providing your responses to the questionnaire after reading the two attached radiology diagnostic accuracy articles implies your consent to participate in this research study. Your participation in this questionnaire would be greatly appreciated. The survey link is provided below.

https://www.surveymonkey.com/r/K6HM7ZW

Yours Truly,

Betty Anne Schwarz
Principal Investigator
Co-Investigator: Dr. Wael Shabana
Middlesex University Supervisor: Dr. Annette Fillery-Travis

Appendix 18: RadSTARD 2 Page Summary and Likert Rating Scale

RadSTARD –Radiology Standards for the Reporting of Diagnostic Accuracy Studies Summary Document with Likert Scoring

1. TITLE & ABSTRACT	Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract. *Use the term 'diagnostic accuracy' in the title of the study(STARD item # 1).*	**Please indicate if this item increased your level of interpretation of the diagnostic accuracy article. 1 – least agree and 10 most agree.** 1 2 3 4 5 6 7 8 9 10
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare the index test with the reference standard for diagnosis of a specific condition.** *The aim is to compare*	1 2 3 4 5 6 7 8 9 10

	the index test to the reference standard (STARD item # 2).	
3. METHODS, PATIENT ELIGIBILITY, DATA COLLECTION	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided. *Specify whether patients studied were receiving primary or secondary care (STARD item # 3).*	1 2 3 4 5 6 7 8 9 10
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.** *Specify if the participants received either the index test or the reference standard (STARD item # 3).*	1 2 3 4 5 6 7 8 9 10
5.	Was patient selection consecutive or non-consecutive? *Specify how participants were enrolled (STARD item # 5).*	1 2 3 4 5 6 7 8 9 10
6.	**Data collection (prospective or retrospective) should be provided with start dates and end dates.** *Record that data was collected as a chart review (retrospectively) or prospective analysis and include the dates the study was done (STARD item # 6 and 14).*	1 2 3 4 5 6 7 8 9 10
7. SAMPLE SIZE	**Sample size and limitations should be provided.** *Providing a sample size reflects the power of the study results (not in the STARD).*	1 2 3 4 5 6 7 8 9 10
8. DIAGNOSTIC TEST METHODS	Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale. *Reporting the reference standard which is the test that is commonly used in medical practice. Rationale for its use should also be defined (STARD item # 7).*	1 2 3 4 5 6 7 8 9 10
9. IMPERFECT REFERENCE STANDARD	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be .justified.** *There is no perfect reference standard – alternatives may consist of a panel standard, composite reference standard, latent class analysis or intrinsic reference standard (not in STARD).*	1 2 3 4 5 6 7 8 9 10
10. TEST	Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies. *A thorough description of the index test and reference standard should be provided so that the reader can interpret if*	1 2 3 4 5 6 7 8 9 10

	the same test could be performed in their institution (STARD item # 8).	
11.	**Sound theoretical physics basis of the index test should be provided for new techniques.** *Description of physics for the index test is very pertinent to radiology to facilitate replication (not in STARD).*	1 2 3 4 5 6 7 8 9 10
12.	**Modifications during the study should be reported if they occurred.** *Describe if the index test or reference standard was modified as this could impact diagnostic accuracy and alter data analysis (not in STARD).*	1 2 3 4 5 6 7 8 9 10
13. ANALYSIS	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests. *It is important to specify which cut-off values were used for a specific diagnosis (similar to STARD item # 9).*	1 2 3 4 5 6 7 8 9 10
14.	The training and number of investigators should be described including **any extra training for new techniques.** *Whether any extra training was required to interpret the index test should be known (similar to STARD item # 10).*	1 2 3 4 5 6 7 8 9 10
15.	Whether readers were blinded to prior test results or clinical information should be known. *The blinding of interpreters is essential to prevent bias (STARD item # 11).*	1 2 3 4 5 6 7 8 9 10
16. RESULTS	Radiology Diagnostic Accuracy Trials should report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations. *The provision of a flow diagram clearly illustrates how many participants were tested or not resulting in final analysis (STARD item # 16).*	1 2 3 4 5 6 7 8 9 10
17.	Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and **concurrent therapies** should be provided. *The provision of concurrent therapies should be provided as they may affect the interpretation of the tests (similar to STARD item # 15).*	1 2 3 4 5 6 7 8 9 10
18.	The severity (spectrum) of the disease entity should be explicitly reported. *The time that a disease was present or if was found by screening could impact how readily it is diagnosed (STARD item # 18).*	1 2 3 4 5 6 7 8 9 10
19.	The time difference between the index test and reference test and details of any other treatments provided between the two tests should be provided. *If there was a delay*	1 2 3 4 5 6 7 8 9 10

	between the conduct of either test this could impact their level of diagnostic accuracy (STARD item # 17).	
20.	Adverse events should be reported for either the index test or reference standard. *It is important to state if there were any incidents of adverse events when conducting either test (STARD item # 20).*	1 2 3 4 5 6 7 8 9 10
21. STATISTICS	Radiology Diagnostic Accuracy Trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI). *Calculated confidence intervals and p-values define how precise the estimates are for the population chosen under study (similar to STARD item # 21).*	1 2 3 4 5 6 7 8 9 10
22.	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers for the index test **and reference standard**; and describe how this data was handled. *All indeterminate results from the index test and the reference standard should be reported as ignoring such results can bias measures of diagnostic accuracy (similar to STARD item # 22).*	1 2 3 4 5 6 7 8 9 10
23.	Radiology Diagnostic Accuracy Trials should include estimates of **inter-observer** variability **including > 3 observers with variable expertise.** *A description of analysis that is planned should be described apriori with > 3 observers of variable expertise as this will decrease inter-observer variability (similar to STARD item # 23).*	1 2 3 4 5 6 7 8 9 10
24.	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility. *This is done by stating the level of Kappa where the statistic impacts the level of intra and inter-observer variability (STARD item # 24).*	1 2 3 4 5 6 7 8 9 10
25. DISCUSSION	The clinical relevance of the study findings should be provided. *Discussion could include comparing and contrasting to previous studies plus describing any limitations to the current study (STARD item # 25).*	1 2 3 4 5 6 7 8 9 10

Appendix 19: Questionnaire for Evaluation of the RadSTARD

Content of the Tool

1. Did the RadSTARD cover all important items? Yes______ No ______

Please comment if additional items should be included:

2. Were any RadSTARD items omitted, added or modified? Yes___ No_____

Please provide any additional comments:

Elaboration Document

3. Was the elaboration document easy to understand? Yes ___ No____

Please provide any additional comments:

Quality of Reporting

4. Did you find the RadSTARD useful in rating the quality of the reporting for the diagnostic accuracy article?
Yes___ No____

Confidence Level

5. Did the provision of the RadSTARD tool increase your level of confidence to interpret the article?
Yes___ No_____

Use Again

6. Would you use the RadSTARD again? Yes___ No____

Please provide any further comments or suggestions you think the RadSTARD requires:

Thank you!

Appendix 20: Reminder Emails

Email to Radiology Resident

Hi Mac,

In follow-up to my previous emails I was wondering if you had a chance to complete the survey that I sent out to the residents re: RadSTARD study?

If you are willing, could you kindly review the diagnostic accuracy article by Henes et al. (2012) and rate your level of confidence in interpreting the article utilizing the RadSTARD tool and accompanying Elaboration Document.

The link to access the survey (to rate your confidence level) is provided below. All your responses are anonymous.

https://www.surveymonkey.com/r/5WK2VNP

Please contact me should you have any questions.

Thanks very much,

Betty Anne

Appendix 21: Regulatory Approval Documents

Ottawa Health Science Network Research Ethics Board/Reseau des sciences de la sante d'Ottawa Conseil d'ethique de la recherche

Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax : 613-761-4311

April 11, 2014

Ms. Betty Anne
Schwarz Radiological
Sciences

Ottawa Hospital - Civic Campus

751 Parkdale, Suite
502 Ottawa, ON

K1Y 1J7

Dear Ms. Schwarz:

Re: Protocol # 20140142-01H 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials'

Protocol approval valid until - April 10, 2015

Thank you for the email of April 11, 2014. I am pleased to inform you that this protocol underwent delegated review by the Ottawa Health Science Network Research Ethics Board (OHSN-REB) and is approved for recruitment of English speaking participants only. This approval allows the Principal Investigator listed on the application to invite eight radiologists to participate in the needs assessment and Delphi technique that is required to develop the revised STARD-DI checklist. No changes, amendments or addenda may be made to the protocol or the consent form without the OHSN-REB's review and approval.

Your request for French exemption has been granted; the study may proceed in English only.

Approval is for the following:

- Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials Protocol version dated April 11, 2014

- English Email Invitation to Radiologists (Version 2) dated April 10, 2014

- English Email Invitation to Radiological Experts Who Agree to Participate (Version 2) dated April 10, 2014 - English STARD-DI (Version 2) dated April 4, 2014

The REB no longer requires a 'valid until' date at the bottom of all approved informed consent forms. The consent forms currently approved for use by the REB are listed above.

Please note: This approval allows the Principal Investigator listed on the application to invite eight radiologists to participate in the needs assessment and Delphi technique that is required to develop the revised STARD-DI checklist. Once the participants deem that a revised tool (STARD-DI) is required, a Protocol Amendment Report will need to be submitted with a copy of the revised STARD-DI. This Protocol Amendment Report and Version 3 of your tool MUST receive approval from the OHSN-REB before you are able to pilot the tool with the Residents and Fellows of the Department of Medical Imaging.

If the study is to continue beyond the expiry date noted above, a Renewal Form should be submitted to the REB approximately six weeks prior to the current expiry date. If the study has been completed by this date, a

Termination Report should be submitted. The Ottawa Health Science Network Research Ethics Board (OHSN-REB) was created by the merger of both the Ottawa Hospital Research Ethics Board (OHREB) and the Human Research Ethics Board (HREB) for meetings held at the University of Ottawa Heart Institute.

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; the International Conference on Harmonization - Good Clinical Practice: Consolidated Guideline; and the provisions of the Personal Health Information Protection Act 2004.

Yours sincerely,

Raphael Saginur, M.D.

Chairperson

Ottawa Health Science Network Research Ethics Board
RS/kh

Ottawa Health Science Network Research Ethics Board/ Conseil d'ethique de la recherche du Roseau de science de la sante d'Ottawa

Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax : 613-761-4311

October 23, 2014

Ms. Betty Anne Schwarz

Radiological Sciences

Ottawa Hospital - Civic Campus
751 Parkdale, Suite 502

Ottawa, ON

K1Y 1J7

Dear Ms. Schwarz:

Re: Protocol # 20140142-01H 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials'

Thank you for the email October 22, 2014. The Protocol Amendment Report dated September 29, 2014 is approved.

Approval of this amendment includes the following:

–Revised Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials Protocol version dated September 28, 2014
- Revised English Email Invitation to Radiologists (Version 3) dated October 16, 2014

–Revised English STARD-DI (Version 3) dated October 22, 2014

Please note: This approval allows the Principal Investigator listed on the application to invite eight radiologists to participate in the needs assessment and Delphi technique that is required to develop the revised STARD-DI checklist. Once the participants deem that a revised tool (STARD-DI) is required, a Protocol Amendment Report will need to be submitted with a copy of the revised STARD-DI. This Protocol Amendment Report and Version 4 of your tool MUST receive approval from the OHSN-REB before you are able to pilot the tool with the Residents and Fellows of the Department of Medical Imaging (Part 2).

The Ottawa Health Science Network Research Ethics Board (OHSN-REB) was created by the merger of both the Ottawa Hospital Research Ethics Board (OHREB) and the Human Research Ethics Board (HREB) for meetings held at the University of Ottawa Heart Institute.

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; the International Conference on Harmonization - Good Clinical Practice: Consolidated Guideline; and the provisions of the Personal Health Information Protection Act 2004.

Ethical approval remains in effect until April 10, 2015.

Yours sincerely,

Raphael Saginur, M.D.

Chairperson

Ottawa Health Science Network Research Ethics Board
/kh

Ottawa Health Science Network Research Ethics Board/ Conseil d'ethique de la recherche du Reseau de science de la sante d'Ottawa

Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax : 613-761-4311

August 4, 2015

Ms. Betty Anne Schwarz

Radiological Sciences

Ottawa Hospital - Civic Campus

751 Parkdale, Suite 502
Ottawa, ON

K1Y 1J7

Dear Ms. Schwarz:

Re: Protocol # 20140142-01H 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials'

Thank you for the email of August 4, 2015. The Protocol Amendment Report dated June 14, 2015 and the Protocol Amendment Report dated July 13, 2015 are approved.

Approval of these amendments includes the following:

–Revised Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials Protocol version dated July 22, 2015
–Usefulness of RadSTARD Questionnaire English Version dated July 13, 2015
–Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials PowerPoint Presentation, version dated August 4, 2015
–RadSTARD — Reporting Tool and Elaboration Document English version dated July 28, 2015
–RadSTARD Summary Sheet and Likert Rating Scale English version dated July 28, 2015
–RadSTARD 2 Page Summary Document English version dated August 4, 2015
- Appendix 8: Email Invitation to Radiology Residents and Fellows version dated July 29, 2015 -
Appendix 9: Follow-Up Email Invitation to Residents and Fellows version dated July 28, 2015

We also acknowledge receipt of the following documents:

–Email correspondence of June 11, 2015 from Dr. Ryan outlining his support in reference to presenting a 10-15 PowerPoint presentation to the radiology residents and fellows at a Radiology Academic Half Day
–Email correspondence of June 11, 2015 from Dr. Hibbert outlining her support in reference to presenting a 10-15 PowerPoint presentation to the radiology residents and fellows at a Radiology Academic Half Day
'Comparison of diagnostic accuracy of Magnetic Resonance Imaging and Multdetector Computed Tomography in the detection of pelvic fractures' - Radiology Diagnostic Accuracy Article

–'Diagnostic Accuracy of Magnetic Resonance Imaging in Endometrial Cancer' - Radiology Diagnostic Accuracy Article
-STARD Checklist for Reporting of Studies of Diagnostic Accuracy, version dated January 2003 - The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration, version dated January 7, 2003

Ethical approval remains in effect until April 10, 2016.

Before sending Appendix 9: Follow-Up Email Invitation to Residents and Fellows version dated July 28, 2015 to the Radiology Residents and Fellows, please ensure that the second set of PowerPoint presentation slides that will be used during the presentation have been submitted, alongside a Protocol Amendment Report, for OHSN-REB review and that approval has been granted.

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; the International Conference on Harmonization - Good Clinical Practice: Consolidated Guideline; and the provisions of the Personal Health Information Protection Act 2004.

Yours sincerely,

Raphael Saginur, M.D.

Chairperson

Ottawa Health Science Network Research Ethics Board RS/kh

Ottawa Health Science Network Research Ethics Board/ Conseil d'4thique de la recherche du Reseau de science de la sante d'Ottawa

Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax : 613-761-4311

September 16, 2015

Ms. Betty Anne Schwarz

Radiological Sciences

Ottawa Hospital - Civic Campus

751 Parkdale, Suite 502
Ottawa, ON

K1Y 1J7

Dear Ms. Schwarz:

Re: Protocol # 20140142-01H 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials'

Thank you for the letter of August 28, 2015. The Protocol Amendment Report dated August 28, 2015 is approved.

The content of the slides for the 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials Validation of the RadSTARD Interim Data Analysis' PowerPoint Presentation, received on August 28, 2015, is also approved. Once complete, a final copy of the PowerPoint Presentation is required for our study files.

Ethical approval remains in effect until April 10, 2016.

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy

Yours sincerely,

Raphael Saginur, M.D.

Chairperson

Ottawa Health Science Network Research Ethics Board
RS/kh

Ottawa Health Science Network Research Ethics Board/ Conseil dfethique de la recherche duRoseau de science de la sante d'Ottawa

Civic Box 411 725 Parkdale Avenue, Ottawa, Ontario K1Y 4E9 613-798-5555 ext. 14902 Fax : 613-761-4311

March 10, 2015

Ms. Betty Anne Schwarz
Radiological Sciences

Ottawa Hospital - Civic Campus

751 Parkdale, Suite 502
Ottawa, ON

K1Y 1J7

Dear Ms. Schwarz:

RE: Protocol# - 20140142-01H 'Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials'

Renewal Expiry Date - April 10, 2016

Thank you for the email of March 10, 2015. I am pleased to inform you that your Annual Renewal Request was reviewed by the Ottawa Health Science Network Research Ethics Board (OHSN-REB) and is approved. No changes, amendments or addenda may be made in the protocol without the OHSN-REB's review and approval.

Renewal is valid for a period of one year. Approximately one month prior to that time, a single renewal form should be sent to the REB office.

We acknowledge that recruitment is closed for Phase 1 of your study. Our file has been updated accordingly.

Please note: This approval allows the Principal Investigator listed on the application to continue working with the eight radiologists that have agreed to participate in the needs assessment and Delphi technique that is required to develop the revised STARD-DI checklist. Once the participants deem that a revised tool (STARD-DI) is required, a Protocol Amendment Report will need to be submitted with a copy of the revised STARD-DI. This Protocol Amendment Report and Version 4 of your tool MUST receive approval from the OHSN-REB before you are able to pilot the tool with the Residents and Fellows of the Department of Medical Imaging (Part 2).

OHSN-REB complies with the membership requirements and operates in compliance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans; the International Conference on Harmonization - Good Clinical Practice: Consolidated Guideline; and the provisions of the Personal Health Information Protection Act 2004.

The Tri-Council Policy Statement requires a greater involvement of the OHSN-REB in studies over the course of their execution. As well, you must inform the Board of adverse events encountered during the study, here or elsewhere, or of significant new information which becomes available after the Board review, either of which may impinge on the ethics of continuing the study. The OHSN-REB will review the new information to determine if the protocol should be modified, discontinued, or should continue as originally approved.

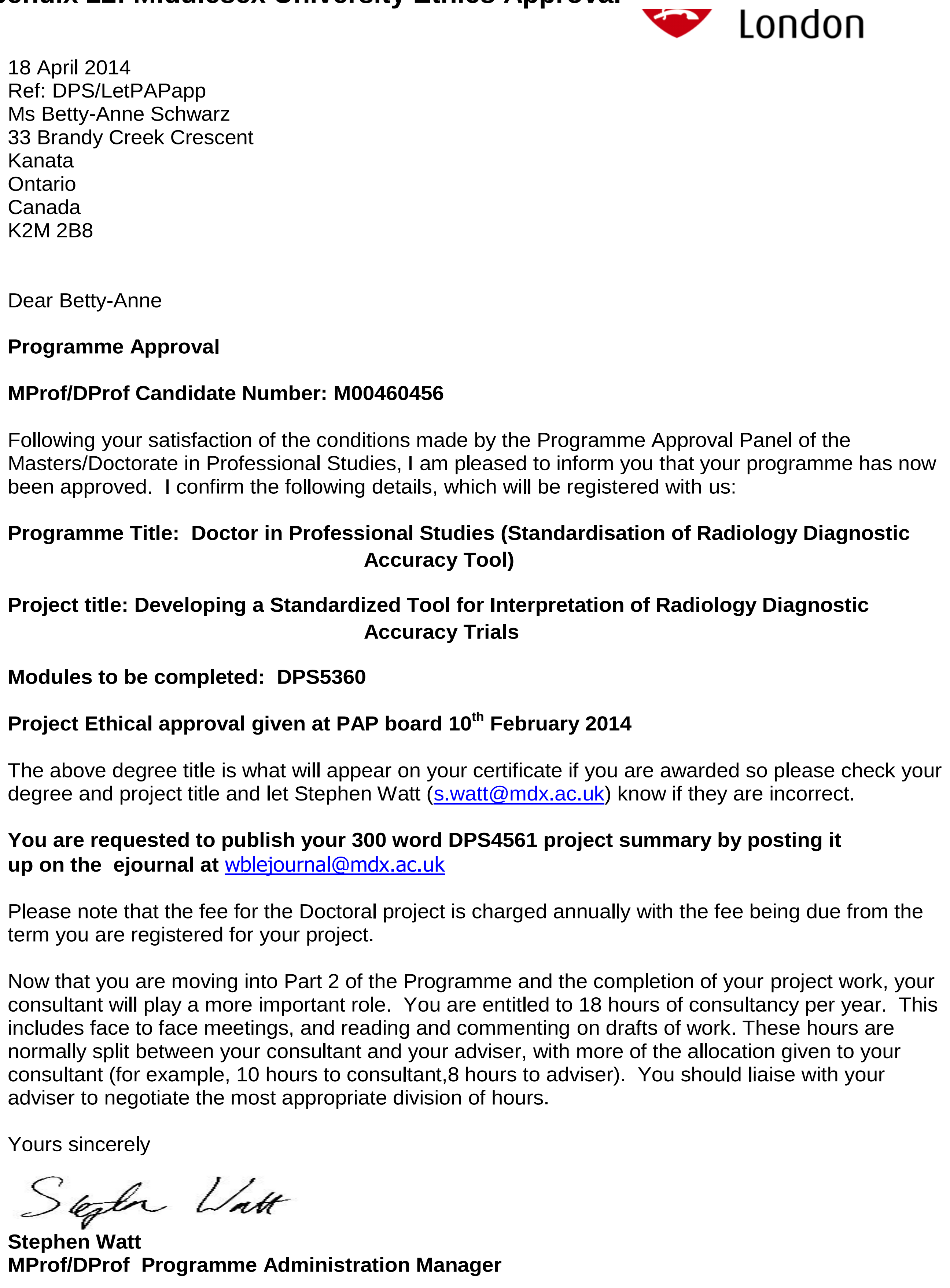
Yours sincerely,

Raphael Saginur, M.D.

Chairperson

Ottawa Health Science Network Research Ethics
Board RS/kh

Appendix 22: Middlesex University Ethics Approval



18 April 2014
Ref: DPS/LetPAPapp
Ms Betty-Anne Schwarz
33 Brandy Creek Crescent
Kanata
Ontario
Canada
K2M 2B8

Dear Betty-Anne

Programme Approval

MProf/DProf Candidate Number: M00460456

Following your satisfaction of the conditions made by the Programme Approval Panel of the Masters/Doctorate in Professional Studies, I am pleased to inform you that your programme has now been approved. I confirm the following details, which will be registered with us:

Programme Title: Doctor in Professional Studies (Standardisation of Radiology Diagnostic Accuracy Tool)

Project title: Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials

Modules to be completed: DPS5360

Project Ethical approval given at PAP board 10th February 2014

The above degree title is what will appear on your certificate if you are awarded so please check your degree and project title and let Stephen Watt (s.watt@mdx.ac.uk) know if they are incorrect.

You are requested to publish your 300 word DPS4561 project summary by posting it up on the ejournal at wblejournal@mdx.ac.uk

Please note that the fee for the Doctoral project is charged annually with the fee being due from the term you are registered for your project.

Now that you are moving into Part 2 of the Programme and the completion of your project work, your consultant will play a more important role. You are entitled to 18 hours of consultancy per year. This includes face to face meetings, and reading and commenting on drafts of work. These hours are normally split between your consultant and your adviser, with more of the allocation given to your consultant (for example, 10 hours to consultant,8 hours to adviser). You should liaise with your adviser to negotiate the most appropriate division of hours.

Yours sincerely

Stephen Watt

Stephen Watt
MProf/DProf Programme Administration Manager

Appendix 23: Email regarding Intellectual Property

Email re: Intellectual Property

Dear Betty Anne,

As discussed, OHRI is of the opinion that any intellectual property developed as a result of your doctoral studies does not fall within the scope of the IP to which OHRI retains ownership by policy. Therefore, we claim no ownership of such IP and as far as we are concerned, there are no IP issues.

I believe that by adding the statement you suggested it will indeed clarify ownership of data with the participants.

Do not hesitate to contact me should you have any question.

Regards,

Anouk

Anouk Fortin, PhD RTTP
Technology Transfer Office
Ottawa Hospital Research Institute
4th floor TOHCC room C4406-b
501 Smyth Road Box 411
Ottawa, Ontario
K1H 8L6
Phone 613-737-8899 x 78930
Fax 613-737-8803
afortin@ohri.ca
www.ohri.ca

From: Schwarz, Betty Anne
Sent: Wednesday, December 11, 2013 12:59 PM
To: Fortin, Anouk
Subject: Intellectual Property Ownership - Doctoral Proposal Middlesex University

Dear Anouk,

Thank you so much for your call back and providing me information with respect to IP. I am a Doctoral Candidate studying at Middlesex University. I am employed by the OHRI and my role is Research Program Manager for the Department of Medical Imaging. In completing my Research Ethics Form (REf) to be attached to my research proposal as required by the university I have been asked the following questions:

Have you made yourself aware of intellectual property issues
Have you clarified with participants the ownership of data

As described, my doctoral project will consist of an action research proposal to develop a revised systematic tool (STARD – Standards for Reporting of Diagnostic Accuracy) that is specific to radiology when interpreting diagnostic accuracy literature. This tool will be developed in Delphi fashion with certain radiologists who are accustomed to research and are experts in their field. Once the tool is developed it will be piloted with the residents and Fellows. For those who agree to participate they will be asked to answer a questionnaire post reading the assigned diagnostic accuracy article to determine if they found the tool to be beneficial. Their responses will be anonymized as they are not required to sign the consent.

With respect to ownership of the data collected from the questionnaire I will insert the comment you suggested to the consent. The wording will be “Data captured from this project will be owned by the investigator”.

Can you kindly advice.

Thanks again,

Betty Anne

Betty Anne Schwarz MSc, BA, RN, Doctoral Candidate
Medical Imaging Research Program Manager
751 Parkdale Ave.Suite 502
Ottawa, ON K1Y 1J7
(613) 798-5555 ext. 17522

Appendix 24: Email Correspondence with Dr. Bossuyt

Email Correspondence with Dr. Bossuyt

April 15, 2014

You are obviously free to create new tools, Betty Anne. My point was about possible confusion around STARD.

Still, I am somewhat surprised that you take on accuracy studies. What prompted you to look at such studies in imaging? Your findings may be relevant for our update of STARD.

I do not want to change your plans, but there are also other areas of imaging research maybe even more in need of reporting guidelines. One such area is the field of reader (agreement) studies; several people have suggested to me that this type of research is in need of better guidelines. Is this something you also came across?

Patrick

Patrick Bossuyt
AMC - University of Amsterdam

From: Schwarz, Betty Anne [mailto:baschwarz@toh.on.ca]
Sent: Monday, 14 April, 2014 17:39
To: P.M.M. Bossuyt
Subject: Re: Permission to Amend STARD checklist

Hi Patrick,

Thank you for elaborating further. What I am proposing to do is work with a group of radiological experts to first do a needs assessment of what items they think should be included when interpreting diagnostic accuracy specific to radiology. Then in Delphi fashion I was going to send them questions with a list of items from both the STARD and the TIDier checklist (interventional tool) to create a new list that can be used by radiologists, residents and Fellows when reading and interpreting radiological diagnostic accuracy tools.

Once we create the new list or tool I would also write up a guidance tool. Then I would test the tool with the radiology residents and Fellows by asking them to read two radiology diagnostic accuracy articles with the both the current STARD and guidance document and the newly created tool with its guidance document to see if they thought there was a difference. For working purposes only the new tool is called STARD-DI (just to do start the first round of the Delphi) but this would not be the name for the new tool when we finished our rounds/cycles.

Are you ok with this?

I work in Ottawa and when I was working on my proposal I discussed it with Dr David Moher who suggested to me that I contact Dr. Hoffman as she had created the TIDier list. As it had to do with interventional procedures he thought it may be relevant. She provided me her list prior to publication and was most interested in providing assistance along the way.

In summary I am proposing to create a new tool based on the results from our needs assessment, studying the current STARD and the TIDier combined and then test it with the trainees as described above.

Please advise.

Thank you,

Betty Anne

Sent from my iPad

On 2014-04-14, at 2:26 AM, "P.M.M. Bossuyt" <p.m.bossuyt@amc.uva.nl> wrote:
Yes, I do have concerns.

When we developed the STARD checklist, we wanted to develop a single checklist that applies to all test accuracy studies, for all types of medical tests. No separate lists for genetics, lab tests, imaging, pathology, etc.

In the past 10 years, that has helped us in the dissemination and implementation of STARD. The same STARD checklist was published in various fields, such as Clinical Chemistry, and Radiology.

So, you are not supposed to change the list, or make amendments for specific types of tests.

What you can do, is help in the interpretation of the items for specific fields, or make instruction tools. Several papers have explained the relevance of STARD for, for example, physical examination (Dave Simel), and for markers in dementia (not yet published, but accepted).

You may also be interested to hear that we are updating the STARD checklist this year.

Kind regards,

Patrick MM Bossuyt, PhD
Professor of Clinical Epidemiology | Dept. Clinical Epidemiology, Biostatistics & Bioinformatics | Academic Medical Center - University of Amsterdam
Room J2-127; PO Box 22700; 1100 DE Amsterdam; the Netherlands | p.m.bossuyt@amc.uva.nl +31(20)566 3240 (voice) +31(20)691 2683(fax)

From: Schwarz, Betty Anne [mailto:baschwarz@toh.on.ca]
Sent: Friday, 11 April, 2014 22:19
To: P.M.M. Bossuyt
Subject: Permission to Amend STARD checklist

Dear Dr. Bossuyt,

Is the STARD tool copyrighted? I am Doctoral student from Middlesex University (distance education) and I work at The Ottawa Hospital in the Department of Radiology as the Research Program Manager. I am proposing to study the STARD tool to determine if amendments are required so that it is more specific for use when interpreting the quality of reporting radiology diagnostic accuracy trials.

Do you have any concerns?

Thank you in advance.

Respectfully,

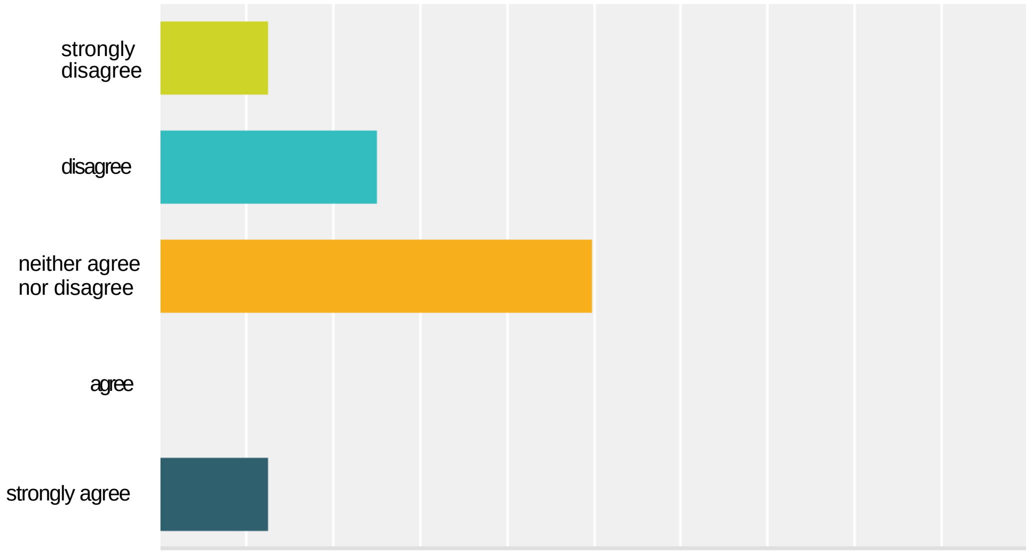
Betty Anne

Betty Anne Schwarz MSc, BA, RN, Doctoral Candidate
Medical Imaging Research Program Manager
502- 751 Parkdale Ave
Ottawa, ON K1Y 1J7
(613) 798-5555 ext. 17522

Appendix 25: Needs Assessment

Needs Assessment - Radiology Diagnostic Accuracy

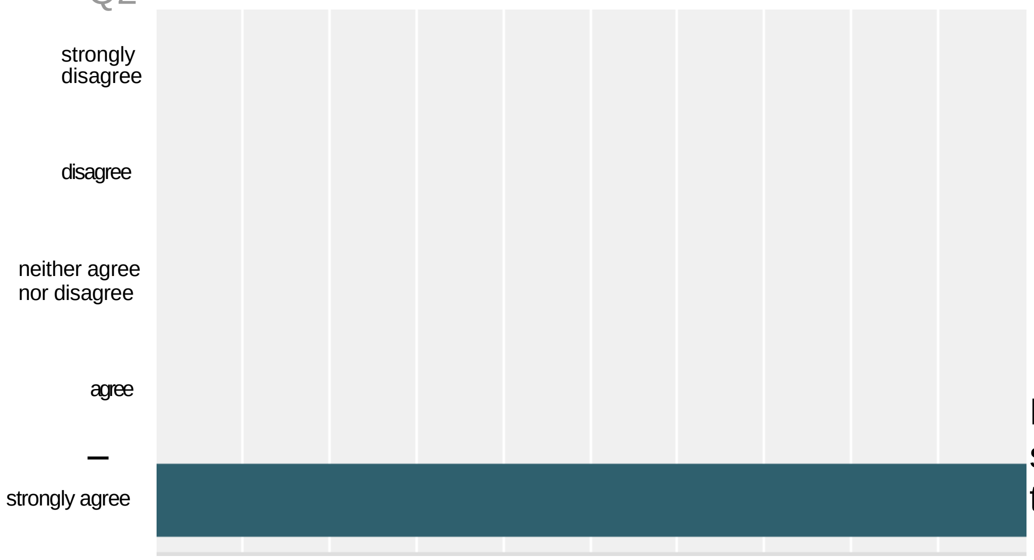
Q1 Should diagnostic accuracy trials mention ‘sensitivity and specificity’ in the title of the study?



Answered: 8 Skipped: 0

Answer Choices	Responses	
strongly disagree	12.50%	1
disagree	25.00%	2
neither agree nor disagree	50.00%	4
agree	0.00%	0
strongly agree	12.50%	1
Total Respondents: 8		

Q2



Eligibility should the

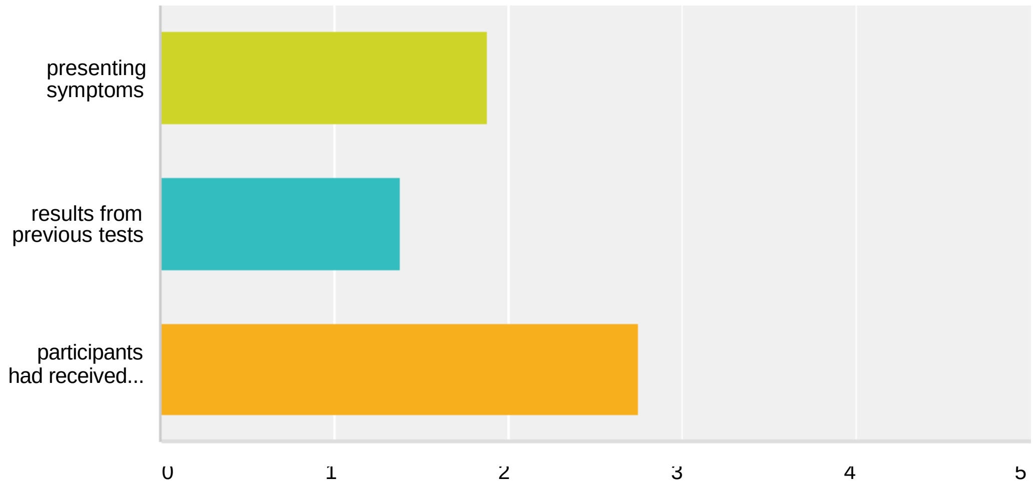
inclusion/exclusion criteria for enrolling the participants be included?

Answered: 8 Skipped: 0

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Answer Choices	Responses	
strongly disagree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	0.00%	0
strongly agree	100.00%	8
Total Respondents: 8		

Q3 Recruitment of participants into the trial. Please rate from 1-3 which items you deem most essential?

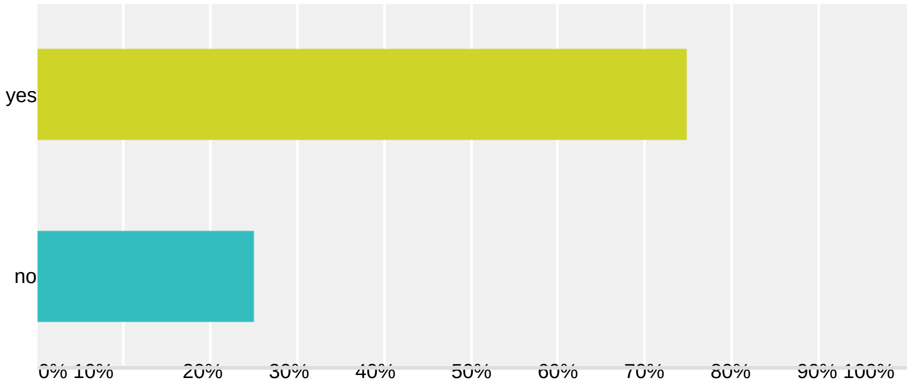


Answered: 8 Skipped: 0

	1	2	3	Total	Average Ranking
presenting symptoms	25.00% 2	37.50% 3	37.50% 3	8	1.88
results from previous tests	0.00% 0	37.50% 3	62.50% 5	8	1.38
participants had received either the index test or the reference standard	75.00% 6	25.00% 2	0.00% 0	8	2.75

Q4 Participant Sampling - Does knowing whether the study population was tested consecutively relevant to your interpretation of diagnostic accuracy studies?

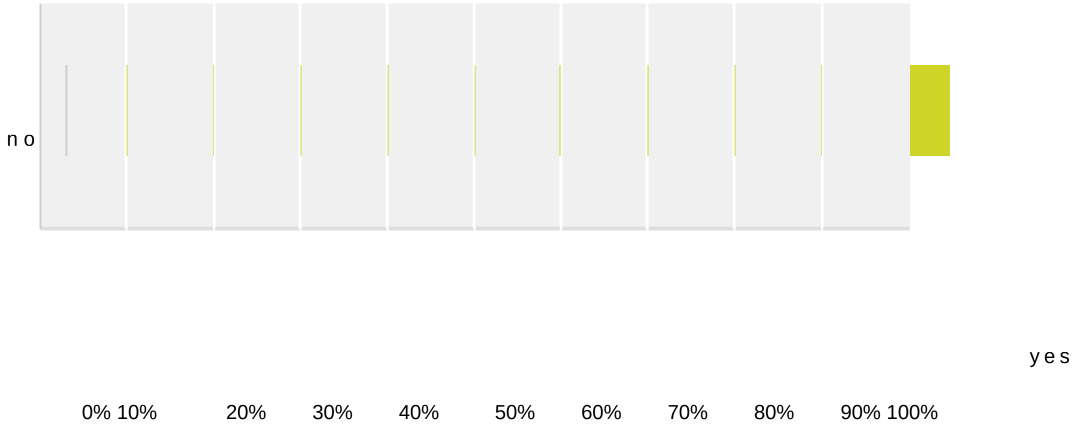
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	75.00%	6
no	25.00%	2
Total		8

Q5 The rationale for use of the reference standard should be stated.

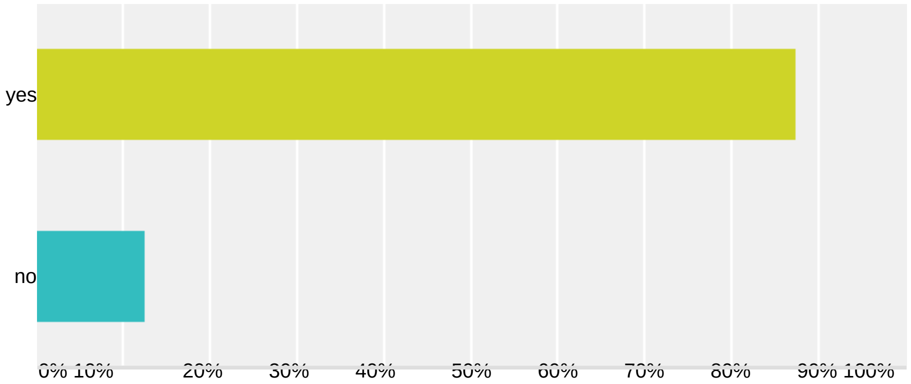
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q6 Make and model of diagnostic imaging equipment should be provided.

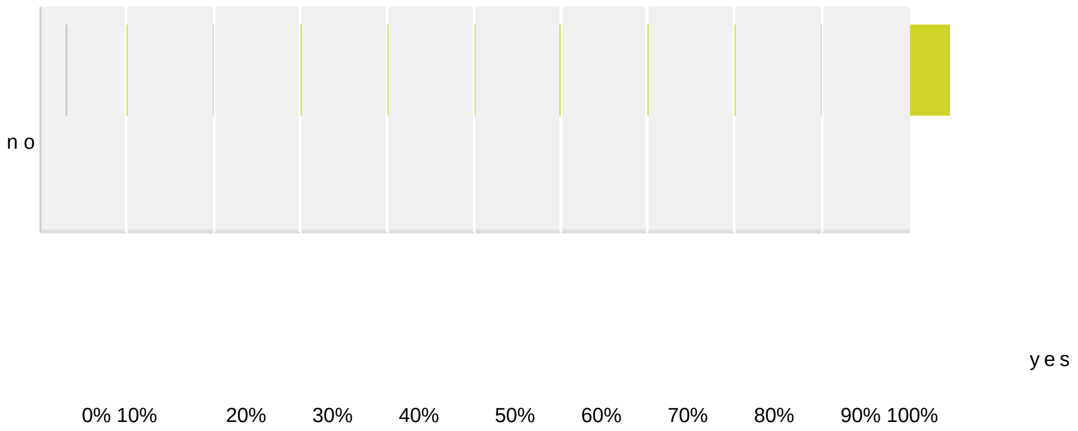
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	87.50%	7
no	12.50%	1
Total		8

Q7 How the test was delivered/performed should be provided.

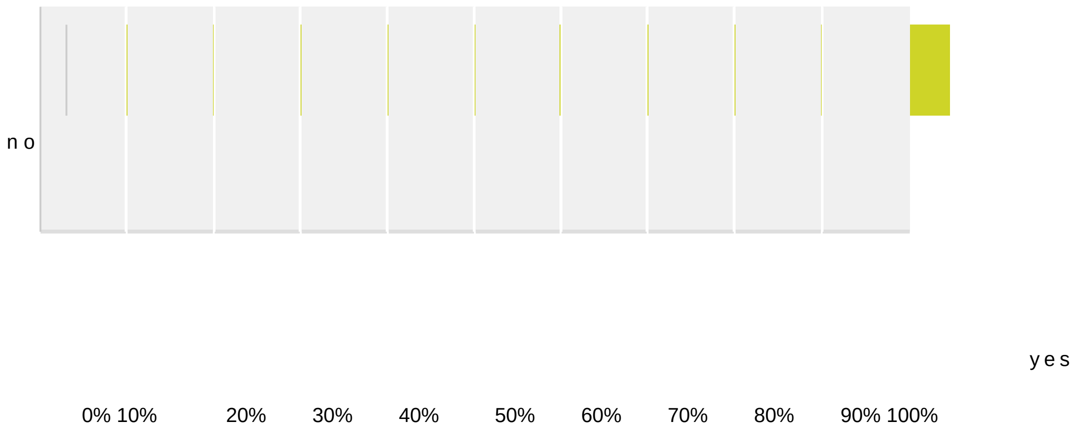
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q8 Definitions for cut-off ranges should be provided.

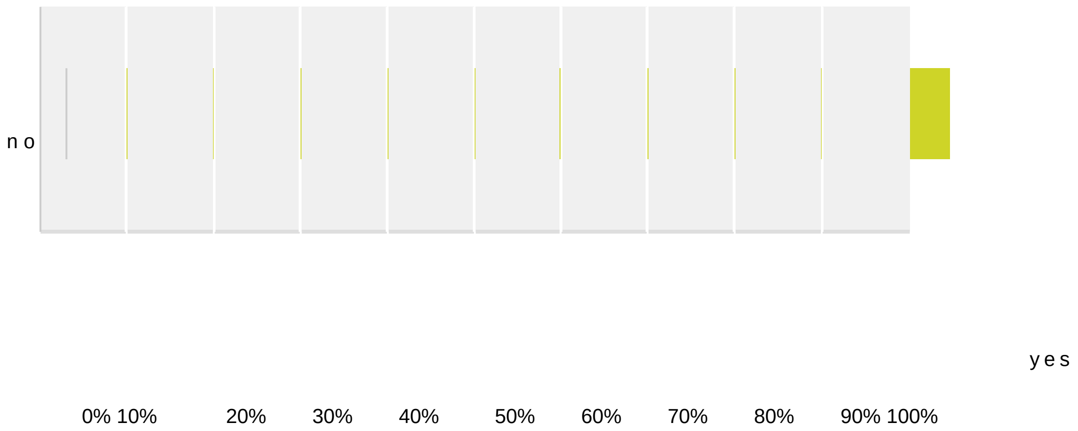
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q9 Training and expertise of the person interpreting the index test and reference standard should be provided.

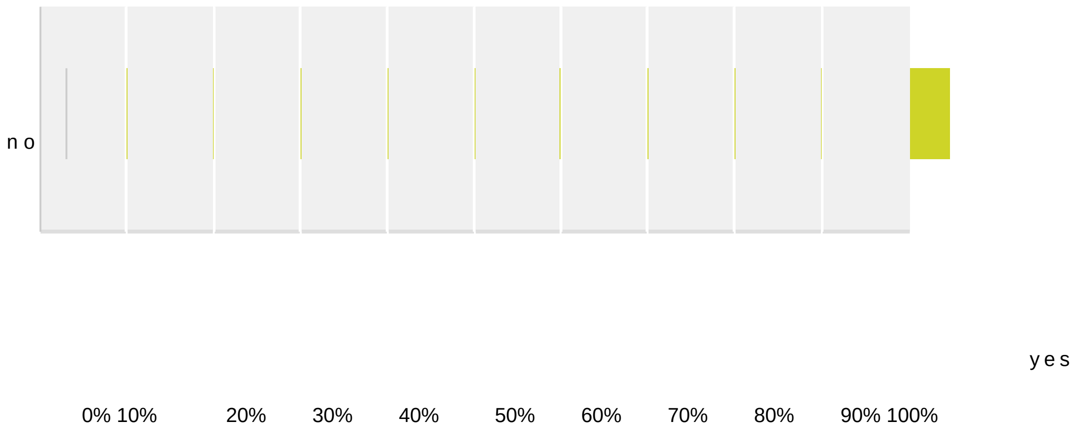
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q10 The interpreter should be blinded to results of previous tests.

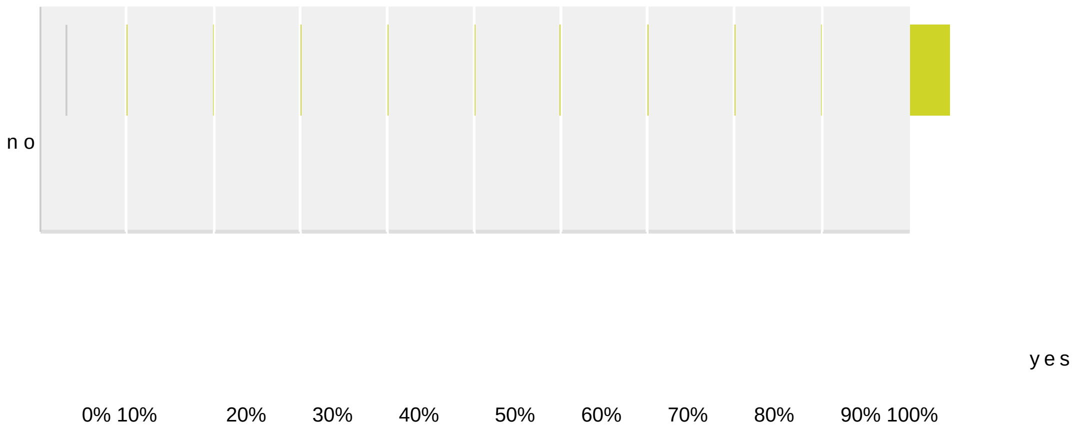
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q11 The methods for calculating diagnostic accuracy between the index test and reference standard should include confidence intervals to quantify uncertainty.

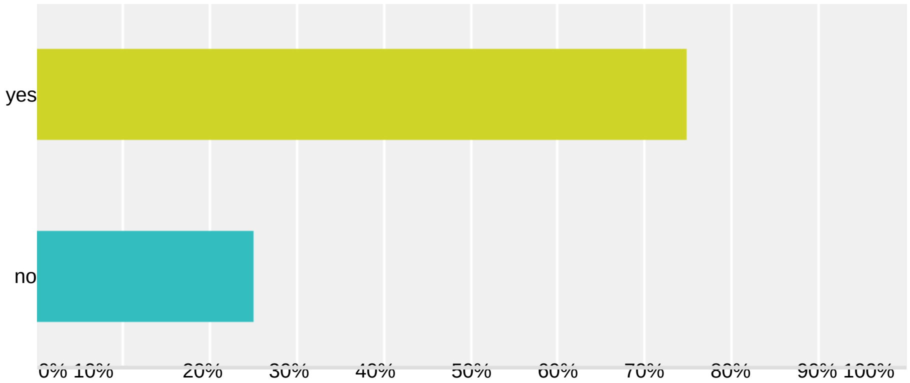
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q12 Should test reproducibility be provided?

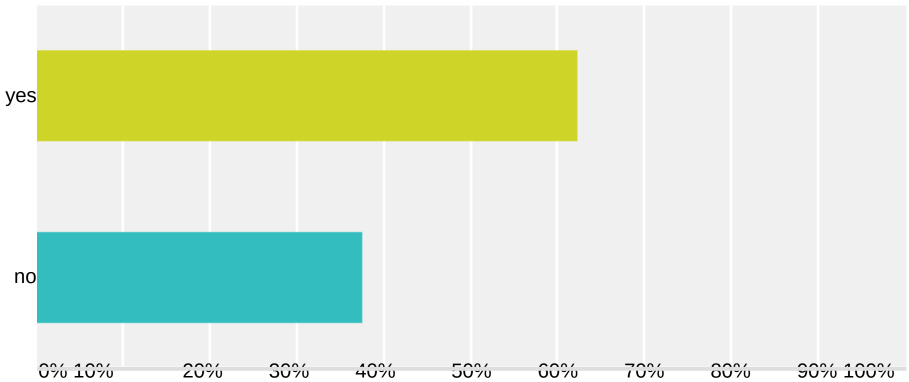
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	75.00%	6
no	25.00%	2
Total		8

Q13 Is a flow diagram describing the number of participants who met elgibility criteria and went on to receive the index test and reference standard required?

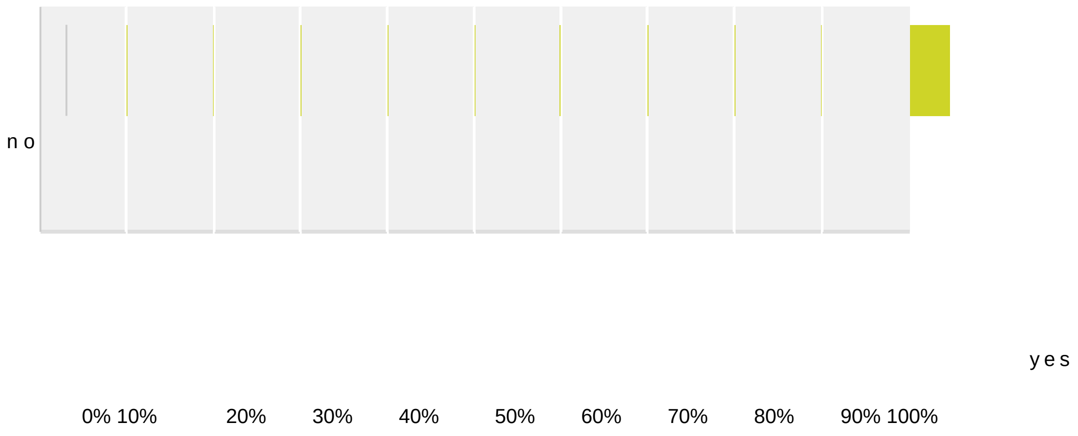
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	62.50%	5
no	37.50%	3
Total		8

Q14 The time interval between the index standard and reference standard should be provided.

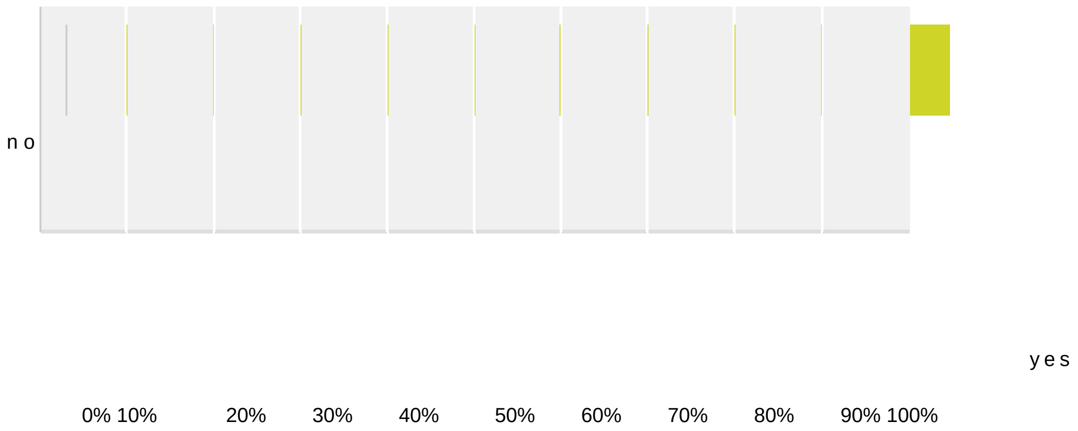
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q15 Should a description of the those with the target disease be provided?

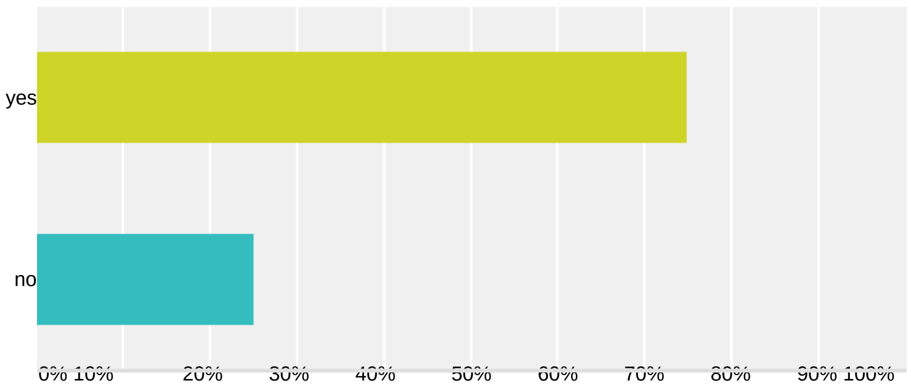
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q16 Should other diagnoses or comorbidites be provided?

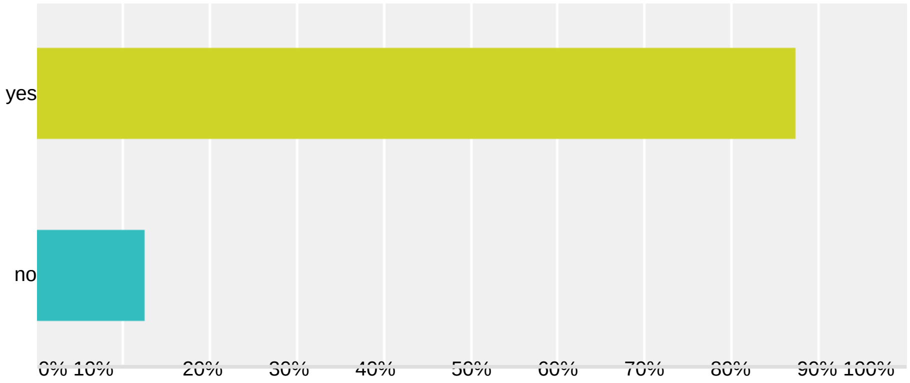
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	75.00%	6
no	25.00%	2
Total		8

Q17 Should a cross tabulation of the results for the index test and reference standard including indeterminate and missing results be provided?

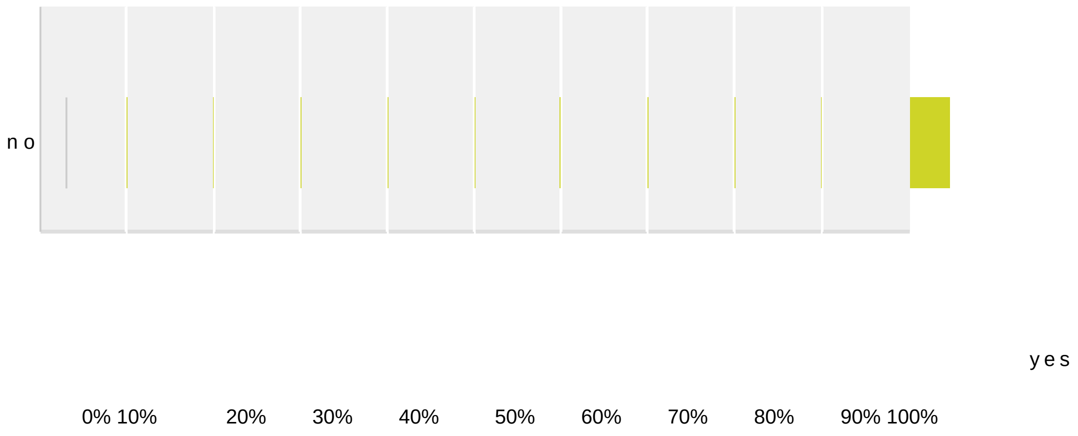
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	87.50%	7
no	12.50%	1
Total		8

Needs Assessment - Radiology Diagnostic Accuracy Q18 Should adverse events be reported?

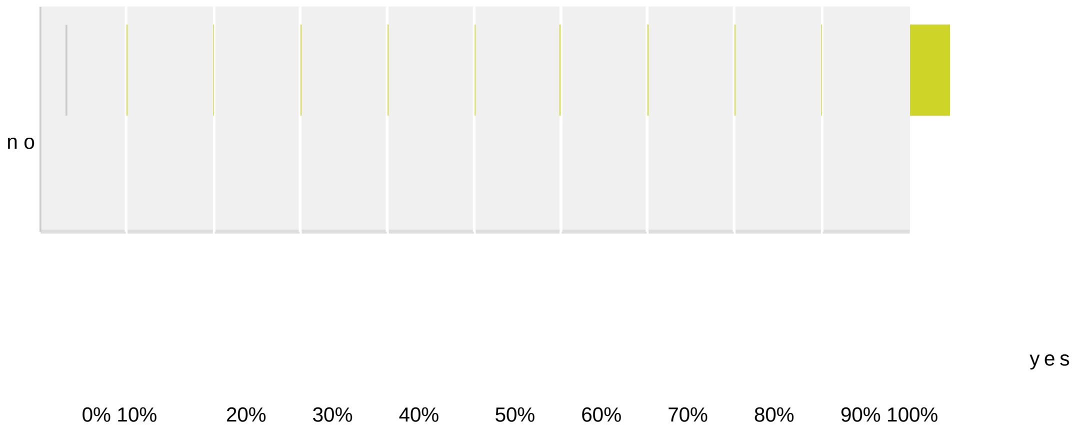
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q19 Confidence intervals 95% are required to describe estimates of diagnostic accuracy and measures of statistical uncertainty.

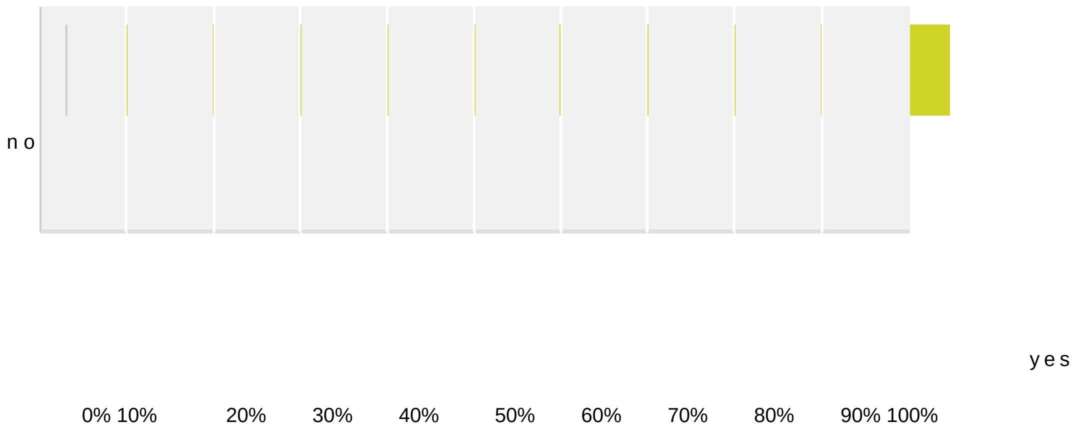
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q20 Should indeterminate results be provided?

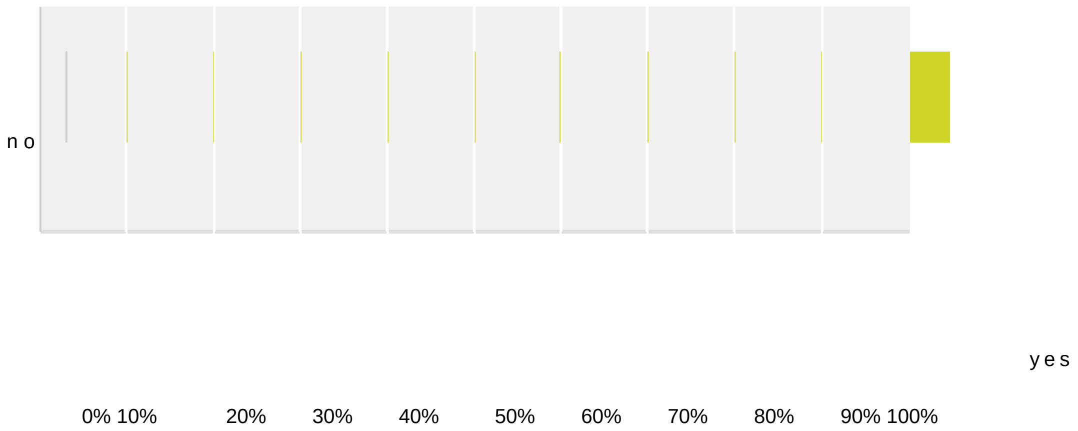
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q21 Should estimates of variability between subgroups of participants be provided if done?

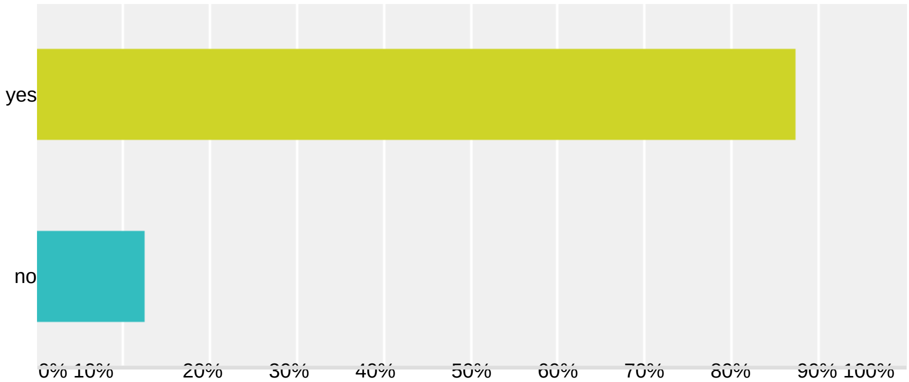
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q22 Should estimates of test reproducibility be provided if done?

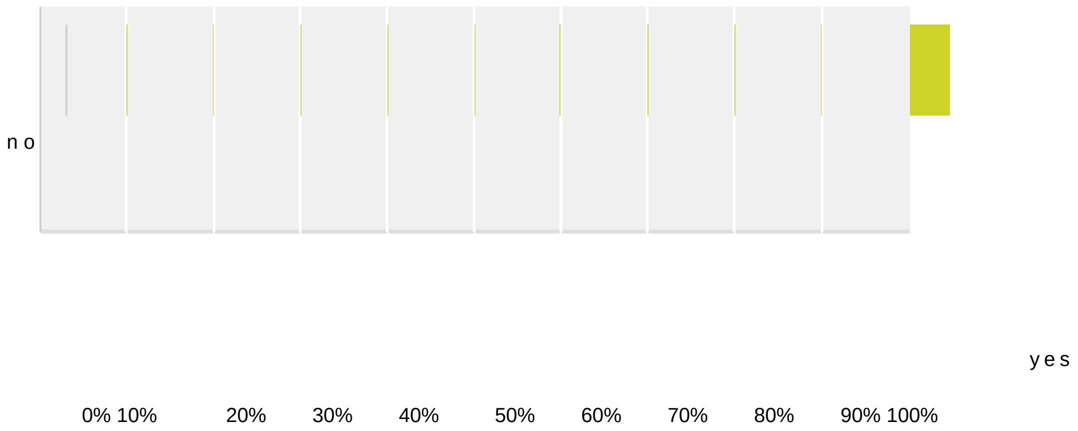
Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	87.50%	7
no	12.50%	1
Total		8

Q23 The clinical applicability of the study findings should be provided.

Answered: 8 Skipped: 0



Answer Choices	Responses	
yes	100.00%	8
no	0.00%	0
Total		8

Q24 Please indicate which additional items you think should also be included when interpreting diagnostic accuracy trials specific to radiology.

Answered: 4 Skipped: 4

#	Responses	Date
1	What part of the result is more generalizable for other vendor equipments Flow chart for practical implementation and quality control	7/31/2014 9:59 PM
2	Precision of the test- interobserver agreement;comment on the sample size and practical limitations; if readers were blinded and if there was training session	7/3/2014 8:02 PM
3	Differences in outcomes. This is the main objective of these types of studies. Was the time different? Could contrast have been avoided? Did the diagnosis change? Was less radiation used? etc...	7/3/2014 11:18 AM

4	Number of patients who had the reference standard test after the index test Inter and intraobserver variability	6/22/2014 11:19 AM

Appendix 26: Emails with Critical Friends

The following email transcripts were critical to the development of the revised tool in relation to my study:

Appendix 25a: Email Exchange with my Mentor regarding Critical Friend #1 (October 30, 2013)

Hi Betty Anne,

I have contacted Dr. Hiremath and he kindly accepted to play the role of the critical friend.
I think it will be better if you can organize with him a meeting to discuss the details.
As I told you, he is a very knowledgeable about methodology and will be a great resource for you.
Plus he is a nice man, somewhere high on the scale of really nice people.

Thanks
Wael

Journal Entry: Most of my exchanges with Critical Friend # 1 were conducted in person. In general, he provided input on the development of the first Delphi round and agreed with my choices utilized for round 2 of the Delphi.

Appendix 25b: Critical Friend # 2 – Introduction (22 Aug 2014)

Email from myself thanking critical friend for agreeing to collaborate with me

Wael informed me that you are agreeable to be my ‘critical friend’ for my doctoral study. Thank you so much – I am most appreciative.

Please find attached a summary of my Doctoral Project, the results of my Needs Assessment (pdf) and round 1 of the Delphi (RADART) in excel format. When I distribute to the radiologists, it will be sent via Survey Monkey.

The first round of the Delphi is a list of questions that I will be sending to the radiological experts that agreed to participate in the study.

Can you please review the list and amend or make suggestions as you deem fit.

The goal of this proposal is to study the STARD to determine if amendment(s) is/are necessary for radiology diagnostic accuracy trials. I am calling this new tool the RADART (Radiology Diagnostic Accuracy Tool).

I am hoping that the working group can come to a consensus after 2-3 rounds of the Delphi.

With respect to the number of responses for the Likert scale I have read that methodologists have recommend five scale points for a unipolar scale, and seven scale points for a bipolar scale.

What do you think is best?

I look forward to your response and thoughts on Round 1 of the Delphi (RADART) tool.

I am also attaching a copy of the STARD tool for reference.

Thanks very much,

Betty Anne

Appendix 25c: Response from Critical Friend #2 - Review of Delphi Round 1 (23 Aug 2014)

Thanks for your e mail. Unfortunately I am on summer vacation until 4th September. However, I will pick this up on my return to the UK and get back to you
that weekend.

Many thanks for making contact & best wishes

MY RESPONSE TO CRITICAL FRIEND #2 (8 Sept 2014)

Dear Nitin,

I hope you had a wonderful summer vacation. Can you please let me know your thoughts on the draft tool previously sent?

Thanks so much – I really appreciate your opinion and expertise.

Sincerely,
Betty Anne

Appendix 25d: Response from Critical Friend # 2 re: Review of Delphi and Likert (8 Sept 2014)

Many thanks for your message. I actually got the date of my return flight wrong in my prior e mail - In fact, I only returned to the UK on Sunday evening... Or rather early Monday morning...

I am going through the material but some of the literature about Likert style questions and scales is complex; and there doesn't seem to be a good statistical definition of 'consensus' in the literature.

This is going to take a few more days to get my head around, but I do believe that I can make a useful contribution. My apologies for the delay - my vacation was somewhat untimely - but I will be in touch later this week with my comments.

Best wishes

MY RESPONSE TO CRITICAL FRIEND #2 (9 Sept 2014)

No worries at all. I am so pleased you are still willing to review the draft for round one of the Delphi. I really value your opinion and look forward to receiving your comments as time permits.

Kindest Regards,

Betty Anne

Appendix 25e: Response from Critical Friend # 2 re: Review of Delphi Round 1 and Needs Assessment (10 Sept 2014)

Quick update

I am nearly finished reviewing the relevant literature and am now working through your excel sheet.

I'm a little unclear at exactly how some of the questions for Delphi Round 1 have been generated (they are not in the original STARD, nor are they suggestions from Q24 of your needs assessment). I am also unsure as to some aspects of the needs assessment - e.g. on what basis were certain items from the original STARD omitted or included - and also, why was a separate Needs Assessment performed (as opposed to just incorporating it as Round 1 of the Delphi Process with opportunities for the panel to make comments). I presume you have formulated a plan for the Delphi process with established criteria for consensus; the basis on which you will accept or reject statements; and what the expected conduct of each round is planned to be. It would be useful for me to be able to access this information.

I will finish my review tomorrow, and write a more detailed response by the end of the week (latest this weekend if I end up taking call on Thu or Fri).

Thanks for your patience and apologies again for the delay due to my extended vacation.

MY RESPONSE TO CRITICAL FRIEND # 2 re: Review of Delphi Round 1 and Needs Assessment (11 Sept 2014)

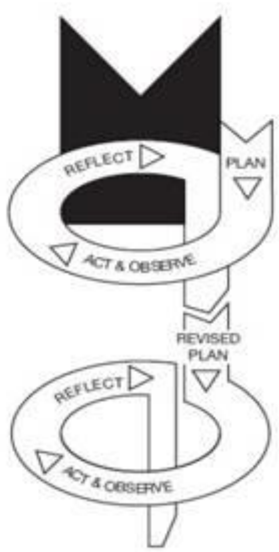
Thank you for your email and preliminary comments. It is greatly appreciated. Please take the time you need – knowing the time difference between us; you were up quite late when you sent this message. I do not want to burden you and I am sincerely grateful for your thoughts and suggestions.

I apologize I didn’t add the proposal before. Please find attached my proposal which explains why I chose to conduct a needs assessment before the Delphi in the methodology section (page 14). Wael and I developed the revised STARD-DI (page 34) after the needs assessment was done – this changed the format of the items for the STARD_DI whereby instead of listing them with a corresponding page number like they are in the STARD, each item is posed as an open end question as is recommended when conducting the Delphi. There is also room for comments for some items.

With respect to formulating a plan for the Delphi process on page 14 it reads: As described by (Hsu & Sandford, 2007) data generated from the Dephi method can consist of both quantitative and qualitative data. Qualitative data is generated from open-ended questions which are generally used with the first iteration of the Delphi. Whereas subsequent iterations are conducted to collect enough data with the goal of reaching a consensus amongst the experts as the tool is built. Due to the mixed methods and flexibility this tool offers is the reason why I am using this data collection tool in this post-positivism paradigm.

Also on page 16 of the proposal it reads: In consultation with Dr. Wael Shabana (local mentor) an initial draft of the amended checklist has been developed whereby elements of the STARD and TIDIER were combined so it can be presented to the radiologists that will be collaborating on this proposal (Appendix 2-B). The initial draft version of the STARD-DI will provide the participant a framework of reference which will be tailored to potentially reduce their concerns regarding their time required to participate in this project.

The radiologists invited to participate in the Delphi will be sent a draft version of STARD-DI and initial list of questions to determine which items they believe needs to be included in the amended STARD – DI checklist (Appendix 2-B). Each round of questions will be provided to the participant via the Delphi technique as this will allow for anonymity. It is anticipated that it will take 4-5 rounds of the Delphi technique to finalize the items that are considered the most relevant in creating this revised checklist. Those participating will be offered two weeks to respond to the questions. Reminder notices will be sent out if their responses are not coming to me in time for analysis before creating the next set of questions to disseminate to the group. Once a consensus has been met with the working group and the revised checklist has been developed it will be piloted within our department to seek validation of the tool.



Kemmis and McTaggart's action research spiral (Koshy et al. 2011)

As per the action research cycle spiral above for Phase 1
Plan – provide the participants (radiological experts) with an initially revised STARD-DI list as a starting point. Ask the experts which items they think should be included in the revised tool
Act + Observe on the process and consequences of the changes (based on their responses)
Reflect on these processes and consequences to develop a revised version of the new tool based on their responses and send back to the experts
Repeat these steps again (replanning, acting and observing, reflecting again and so on…) until the new revised STARD-DI tool has been developed.

Please note that in my original proposal the initial list that would be provided to the radiological experts for the first round of the Delphi was called the STARD_DI. However; so as to not confuse the name of this new tool with the current STARD I am now planning on calling it the RADART – Radiology Diagnostic Accuracy Reporting Tool. I will most likely not change the name of this tool from STARD_DI to the RADART until after it is developed. I am waiting to confirm with Middlesex University how to proceed. Either I submit this change now as an Amendment to the protocol or wait till it's developed (before I test it with the residents and Fellows).

With respect to the initial Needs Assessment, the questions were created by myself and sent to Wael who suggested that I send them to another physician "critical friend". He approved them and I sent the needs assessment out for response. Essentially, many of the items were accepted as being required by the radiological experts that upon reflection speaks to the fact that either the questions could have been worded differently or some of the results rendered plasticity. As they were not a lot of additional comments obtained from the needs assessment we pondered that this may be due to that the radiological experts don't use the STARD much and don't assess the quality of reporting of diagnostic accuracy studies in this manner. Nonetheless, they are keen to participate and in an effort to provide them with enough information for them to evaluate the value of the item when rating the quality of the reporting tool (RADART) and foresee the potential of the tool for others to use when interpreting the literature and increase the quality of reporting radiology diagnostic accuracy studies.

I hope this provides some clarity.

Many thanks again,

Betty Anne

Reference:
Hsu, C. C. & Sanford, B. A. (2007) 'The Delphi Technique: Making Sense of Consensus', Practical Assessment Research & Evaluation, 12 (10), [Online]. Available from: http://pareonline.net/getvn.asp?v=12&n=10 (Accessed: 10 October, 2013).

Koshy, E., Koshy V. & Waterman, H. (2011) ‘What is Action Research’. In: Koshy, E., Koshy V. & Waterman, H. (1st ed.) *Action Research in Healthcare*. London: SAGE, pp.1-24.

Appendix 25f: Response from Critical Friend # 2 re: Methods Employed (15 Sept 2014)

Thank you very much for sending me your project proposal. This was very helpful in gleaning some background to your thesis, and explaining some of the methods. I think it is very well written so far, good job!

I do still have some issues that I am unclear about. These are as follows:-

1. Is RADART a tool assessing Quality of REPORTING, or Quality of STUDY? (i.e. is it the equivalent of STARD or QUADAS?)
- Most of your thesis suggests that it is a measure of quality of reporting - i.e. equivalent of STARD.
- However, there are a few passages (e.g. where you describe the aim of the questionnaire studies for validation), where I became confused, as you mentioned that "Residents/Fellows will be asked to rate the methodological quality of articles" (p20)... this sounds more like QUADAS. I believe there is some overlap (since Quality of reporting does include some element of standards), but I would be grateful if you could clarify this.

2. Methodological Issues
<u>(i) Creation of RADART and Needs Assessment.</u>
- As far as I can surmise, you and Dr. Shabana jointly combined STARD with TIDIER to create RADART v1. This was then sent to the panel with a Needs Assessment. The purpose of the Needs Assessment was to ascertain whether the panel felt that there was a need for RADART at all, or whether STARD was sufficient. Is this correct??

If so, I have 2 comments:

(a) Why combine STARD with unpublished work? There is nothing wrong with this methodology in principle (it may even increase your number of criteria for an initial "long-list", which is favorable). However, I'm not sure it is strictly necessary when established criteria are already in place.

(b) The Needs Assessment showed consensus for the vast majority of the questionnaire. I do not think this is due to reader inexperience. I think it is because most of the questions are relatively incontrovertible methodological points. I was not surprised at the high level of agreement given the content of the questionnaire.
- I was a little surprised that the Needs Assessment did not ask open-ended questions of the panel - e.g. "Do you think STARD is relevant to Radiology?"; "What aspects are least relevant?"; "How can it be tailored to Radiology"; "Is it too long and difficult to use"; etc...
- Either way, I am not really very clear on exactly what the Needs Assessment added in this case. Could it have just been incorporated into Round 1 of Delphi? Anyway I do not think this is a major issue.

<u>(ii) Delphi Process</u>
(a) I'm not clear about the selection process for the panel. As you know, the content of the panel is the foundation of the Delphi process, and relatively strict criteria should be in place. Obviously I know that you have chosen experts from each of the sections (many of whom I know personally and who are excellent researchers) - but I think this needs to be detailed in your thesis.

(b) I am also not clear about the Aim and Plan for each round of the process. Specifically:-
- What is the output of Round 1? Is it information gathering and will you define a checklist from the results? Do you plan to include open-ended questions asking panelists how STARD can be tailored to suit Radiology (and whether it needs to be); how practically usable the checklist is; whether it would be useful to shorten/summarize it; etc...?
- What is the aim of Round 2? I assume you will try to achieve consensus on items which should be included in the formal checklist.
- Rounds 3 and 4 might then be used to narrow down the list. You could consider ranking items in order of importance and having a 5 point "essentials" list - this could also be used in you Questionnaire study for validation.
- Also, How will you define "Consensus" and "Non-consensus"? On what basis will you decide which items to accept, reject, or re-discuss? Will you use a Likert or Dichotomous scale for subsequent rounds?

I am sure you have planned out all of these factors prior to beginning the process, but it would be useful to know your response to these issues.

As for the initial questionnaire, I have the following comments and responses to your queries:-
- 5 or 7 point Likert scale are both acceptable. 5 point Unipolar is fine, and makes data analysis easier. I agree with your methodology.
- I find there is a lot of repetition of material in the questionnaire; some questions of which I am uncertain regarding their utility; some items which have been omitted but were present in STARD; and some questions which are difficult to read/understand.

Final Question:
Out of interest, what did your MSc show? What items in STARD and QUADAS did you find were not applicable to Radiology in general?

Well done again for all your hard work. I recognize that you have a lot of expertise and experience in this area. I am only offering you my opinion based on my reading of the literature and your thesis (which may be incomplete/out-of-date information on both counts). Please feel free to take or leave my suggestions as you see fit.

I have copied Dr. Shabana in to this message to keep him up to date as well.

Many thanks and Best wishes

MY RESPONSE TO CRITICAL FRIEND # 2 re: Methods Employed (15 Sept 2014)

Thank you very much for your suggestions and comments.

Dr. Shabana and I met briefly this afternoon. He has just returned from vacation and is working the late shift. I will review your suggestions about the first round of the Delphi (RADART) with Wael later in the week and get back to you.

In the meantime, please see my responses embedded below:

Kindest Regards,

Betty Anne

I do still have some issues that I am unclear about. These are as follows:-

1. Is RADART a tool assessing Quality of REPORTING, or Quality of STUDY? (i.e. is it the equivalent of STARD or QUADAS?)
- Most of your thesis suggests that it is a measure of quality of reporting - i.e. equivalent of STARD.

- However, there are a few passages (e.g. where you describe the aim of the questionnaire studies for validation), where I became confused, as you mentioned that "Residents/Fellows will be asked to rate the methodological quality of articles" (p20)... this sounds more like QUADAS. I believe there is some overlap (since Quality of reporting does include some element of standards), but I would be grateful if you could clarify this.

Although I thought of calling the new tool the RADART (instead of STARD_DI as it is labeled in the proposal) in discussion with my university Advisor, she thinks it's best if I continue on developing the tool (STARD_DI) and then when it's completed; I'll submit this revised tool to ethics for approval (Ottawa REB) prior to testing it with the residents and Fellows. Then I'll change the name to the RADART with an Amendment to the research ethics board.

To answer your question, this revised/new tool (STARD_DI) [*RADART in the future*] is to provide a tool for radiologists to use when rating the quality of the reporting of diagnostic accuracy studies specific to radiology.

The current STARD tool is not specific in that some items don't pertain to radiology or some items need to be elaborated or some items may need to be added to the current list.

Once this new list is developed, it is hoped that it would benefit radiologists/trainees when interpreting diagnostic accuracy trials specific to radiology. If many of the items in the revised list specific to radiology are not mentioned in radiology diagnostic accuracy studies this could impact one's interpretation of the study leading to biases.

2. Methodological Issues
<u>(i) Creation of RADART and Needs Assessment.</u>
- As far as I can surmise, you and Dr. Shabana jointly combined STARD with TIDIER to create RADART v1. This was then sent to the panel with a Needs Assessment. The purpose of the Needs Assessment was to ascertain whether the panel felt that there was a need for RADART at all, or whether STARD was sufficient. Is this correct??

Not quite. I initially sent out the Needs Assessment to the radiological experts to ascertain what items they thought should remain and what additional items should be included.

Wael and I then met and we created the list based on the results of the needs and the STARD/TiDier). This list is for first round of the Delphi.

If so, I have 2 comments:

(a) Why combine STARD with unpublished work? There is nothing wrong with this methodology in principle (it may even increase your number of criteria for an initial "long-list", which is favourable). However, I'm not sure it is strictly necessary when established criteria are already in place.

I agree with you in principle. The STARD tool is great however as described by (Ochodo et al., 2013).

"The evidence provided reveals that there is still ample room for improvement in the reporting of diagnostic-accuracy studies. With an increase in the awareness of the potential biases in diagnostic-accuracy studies, authors and reviewers of these studies will increasingly use and refer to STARD in their manuscripts.

In light of STARD's impact thus far, what is next? Guidelines are not static, and they need to be revised or extended to accommodate a broader scope or the changing landscape of a research field.

The quality of reporting should also improve if STARD is modified so that it is accompanied by extensions and comparable guidelines for the evaluation of other features of medical tests, and if the STARD website is revamped to make its content more comprehensible to users" (Ochodo et al., 2013:919)

Our modifications/amendments to STARD to create a tool specific to radiology (RADART) will include extensions and/or deleted items along with a comparable guideline.

Reference: Ochodo, E. A. & Bossuyt, P. M. (2013) 'Reporting the Accuracy of Diagnostic Tests: The STARD Initiative 10 Years On', Clinical Chemistry, 59 (6), pp.917-919.

At the time that I was writing my proposal, the TIDieR was not yet published but it is now. Dr. Moher is a Senior Scientist/Methodologist (Methods Center) who developed the CONSORT and has worked with all of the notable leaders in this field of research. He suggested I contact the author of the TIDieR in confidence to ask if she'd share her list with me given that a revised reporting tool specific to radiology has to do with interventions.

I have also contacted Dr. Bossuyt who is the lead developer of the steering committee for the STARD. So he knows what I'm up to. He lifted the copyright of the STARD stating I was free to create a new tool. He asked if I would share what we (radiology) found as our findings may be relevant to future revisions.

(b) The Needs Assessment showed consensus for the vast majority of the questionnaire. I do not think this is due to reader inexperience. I think it is because most of the questions are relatively incontrovertible methodological points. I was not surprised at the high level of agreement given the content of the questionnaire.
- I was a little surprised that the Needs Assessment did not ask open-ended questions of the panel - e.g. "Do you think STARD is relevant to Radiology?"; "What aspects are least relevant?"; "How can it be tailored to Radiology"; "Is it too long and difficult to use"; etc...
- Either way, I am not really very clear on exactly what the Needs Assessment added in this case. Could it have just been incorporated into Round 1 of Delphi? Anyway I do not think this is a major issue.

The questions of the needs assessment could have been worded differently which is something I am very cognizant of for the Delphi.

<u>(ii) Delphi Process</u>
(a) I'm not clear about the selection process for the panel. As you know, the content of the panel is the foundation of the Delphi process, and relatively strict criteria should be in place. Obviously I know that you have chosen experts from each of the sections (many of whom I know personally and who are excellent researchers) - but I think this needs to be detailed in your thesis.

I chose an expert from each body system within radiology as I am proposing to develop a tool to be used by radiologists when reading any radiological diagnostic accuracy study. Hence, I would like their expert opinion to create the tool so that it is universal for all areas of radiology.

(b) I am also not clear about the Aim and Plan for each round of the process. Specifically:-
- What is the output of Round 1? Is it information gathering and will you define a checklist from the results? Do you plan to include open-ended questions asking panelists how STARD can be tailored to suit Radiology (and whether it needs to be); how practically usable the checklist is; whether it would be useful to shorten/summarize it; etc...?
- What is the aim of Round 2? I assume you will try to achieve consensus on items which should be included in the formal checklist.
- Rounds 3 and 4 might then be used to narrow down the list. You could consider ranking items in order of importance and having a 5 point "essentials" list - this could also be used in you Questionnaire study for validation.
- Also, How will you define "Consensus" and "Non-consensus"? On what basis will you decide which items to accept, reject, or rediscuss? Will you use a Likert or Dichotomous scale for subsequent rounds?

I am sure you have planned out all of these factors prior to beginning the process, but it would be useful to know your response to these issues.

The responses to each Delphi round will be analyzed and summarized when each subsequent round is sent to the participants on each successive round. Knowing the responses of their peers (anonymously) may result in them changing their response in the next round of the Delphi which will eventually result in meeting consensus.

for the initial questionnaire, I have the following comments and responses to your queries:-
- 5 or 7 point Likert scale are both acceptable. 5 point Unipolar is fine, and makes data analysis easier. I agree with your methodology. OK – thank you.

- I find there is a lot of repetition of material in the questionnaire; some questions of which I am uncertain regarding their utility; some items which have been omitted but were present in STARD; and some questions which are difficult to read/understand.
I will review this later in the week with Wael and send back a revised list to you for further consideration. Thanks very much for your suggestions, comments etc.

Final Question:
Out of interest, what did your MSc show? What items in STARD and QUADAS did you find were not applicable to Radiology in general?

Findings: Of the 950 citations identified, 36 trials were deemed relevant for data abstraction and analysis. Results from both time intervals were compared. Since the development of the STARD the quality of reporting abdominal imaging diagnostic accuracy trials improved in 16 items (64%) and 3 items (12%) deteriorated. Review of the methodological quality with the QUADAS revealed that 10 items (71%) had improved and 3 items (21%) declined.

Abbreviated Results for STARD:
The following table provides a descriptive analysis of the STARD items that were found to have improved when the pre-STARD (1996-2003) was compared to the post-STARD (2004-2011)

Table 4: Summary of Changes from pre-STARD (1996-2003) to post-STARD (2004-2011)

STARD Item Number	Pre-STARD N=19 (%)	Post-STARD N=17 (%)	(%) Improvement
Item #1	18(95)	17(100)	(5)
Item #7	18(95)	17(100)	(5)
Item #8	13(68)	17(100)	(32)
Item #9	17(89)	16(94)	(5)
Item #11	10(53)	10(59)	(6)
Item #12	13(68)	16(94)	(26)
Item #13	2(11)	8(47)	(36)
Item #14	15(79)	16(94)	(15)
Item #16	11(58)	15(88)	(30)
Item #17	8(42)	8(47)	(5)
Item #18	14(74)	13(76)	(2)
Item #19	5(26)	9(53)	(27)
Item #21	10(53)	13(76)	(23)
Item #22	12(63)	14(82)	(19)
Item #23	2(11)	12(71)	(60)
Item #24	1(5)	7 (41)	(36)

In summary, since the development of the STARD the results from this review between the two time intervals revealed that the quality of reporting diagnostic accuracy trials in patients presenting with abdominal pain to the emergency department improved in 16 (64%) of the 25 STARD recommendations whereas; three (12%) items were found to have declined and 6 items (24%) remained the same. My findings concluded some evidence of adherence to the guidelines and that certain items didn't pertain to radiology or could be refined further.

Within the introduction of my Doctoral Proposal I wrote:

“Numerous reasons have been postulated why there's been such a slow adoption of the STARD statement. It takes time for new guidelines to be adopted by journals and authors. In addition it has been found that journals have given poor instructions to authors on how to incorporate the use of the STARD checklist. Suffice to say there is a gap in the literature in that recommendations are necessary to improve adherence to the guidelines when reporting diagnostic accuracy findings. As described by Ochodo and Bossuyt (2013) guidelines such as the STARD are not static and therefore it's possible that amendments to the guidelines would be beneficial to various subspecialties. Although the CONSORT checklist has been amended twice since its initial publication the STARD checklist has remained the same. Therefore; the goal of this proposal is to study the STARD to determine if amendment(s) is/are necessary for radiology diagnostic accuracy trials to develop a new tool and reporting guideline specific to radiology”.

Well done again for all your hard work. I recognize that you have a lot of expertise and experience in this area. I am only offering you my opinion based on my reading of the literature and your thesis (which may be incomplete/out-of-date information on both counts). Please feel free to take or leave my suggestions as you see fit.
I very much appreciate your comments and suggestions.I have copied Dr. Shabana in to this message to keep him up to date as well.

Many thanks and Best wishes

Thank you very much!
Betty Anne

Appendix 25g: Response from Critical Friend # 2 re: STARD_DI Delphi Round 1 (22 Sept 2014)

Easier to rewrite how I think it could be presented differently.
See attached.

Best wishes

STARD-DI checklist for the first round of the Delphi Technique **APPENDIX 2-B**			
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**
Title & Abstract	1	Should Radiology Diagnostic Accuracy Trials include the words "Sensitivity & Specificity" or "Diagnostic Accuracy" in the Title/Abstract? Please choose which one below.	1 □ 2□ 3□ 4□ 5□
	(i)	"Sensitivity & Specificity" in the Title/Abstract?	1 □ 2□ 3□ 4□ 5□
	(ii)	"Diagnostic Accuracy" in the Title/Abstract?	1 □ 2□ 3□ 4□ 5□
	2	Should Radiology Diagnostic Accuracy Trials explicitly state that the Aim is to compare Index test with Reference standard for diagnosis of a specific condition	1 □ 2□ 3□ 4□ 5□
	3	What should be included in the Title and Abstract?	Please provide comment: ______________
	4	Does this need to be included in a checklist for Radiology Diagnostic Accuracy?	1 □ 2□ 3□ 4□ 5□
METHODS			

Patient selection, recruitment and sampling	5	Should Diagnostic Accuracy Trials provide Inclusion and Exclusion criteria for patients?	1 □ 2□ 3□ 4□ 5□
	6	Should Radiology Diagnostic Accuracy Trials report:-	
		Details of setting and location of study (e.g. whether primary or secondary care) ?	1 □ 2□ 3□ 4□ 5□
	(i)	Whether data collection is prospective or retrospective ?	1 □ 2□ 3□ 4□ 5□
	(ii)	Whether patient selection was consecutive or non-consecutive?	1 □ 2□ 3□ 4□ 5□
	(iii)	Sample size and limitations?	1 □ 2□ 3□ 4□ 5□
	(iv)	Whether a Power calculation was performed?	1 □ 2□ 3□ 4□ 5□
	(v)	Start and End dates of the study?	1 □ 2□ 3□ 4□ 5□
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**
	7	Should a thorough description of the nature of the disease be presented?	1 □ 2□ 3□ 4□ 5□
Diagnostic test methods	8	Should Radiology Diagnostic Accuracy Trials explicitly state the Reference Standard and its rationale?	1 □ 2□ 3□ 4□ 5□
Imperfect Reference Standard	9	If the Reference Standard is unavailable or imperfect, should trials state what alternative was used and justify the alternative? For example: New MRI technique for rotator cuff tear (index test) compared to arthroscopy (reference standard). If only 10% went on to surgery this would create a verification bias. Test only works in patients with severe disease creating a verification bias. Therefore would a mixed standard of reference be required if several orthopedic surgeons think there is no rotator cuff injury?	1 □ 2□ 3□ 4□ 5□
	10	If you do not think a mixed standard or panel standard would be of benefit, please insert your comments of what would be a viable alternative for diagnostic studies with an imperfect reference standard?	Please provide comment: ________________
Test	11	Technical specifications for index and reference tests should be reported:-	
	(i)	In All Diagnostic Accuracy Trials?	1 □ 2□ 3□ 4□ 5□
	(ii)	Only where the Test is not the standard of care?	1 □ 2□ 3□ 4□ 5□
	12	Should Radiology Diagnostic Accuracy trials explicitly state generalisability of the technique to other vendor equipment?	1 □ 2□ 3□ 4□ 5□
	13	A sound theoretical Physics basis of the test should be provided:	
	(i)	For all techniques used?	1 □ 2□ 3□ 4□ 5□
	(ii)	Only for New Techniques?	1 □ 2□ 3□ 4□ 5□

	14	If the test was modified during the study, do you need to know?	1 □ 2□ 3□ 4□ 5□
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**
Analysis	15	Should Radiology Diagnostic Accuracy Trials report Cut-off values or specific diagnostic criteria for Index and Reference tests?	1 □ 2□ 3□ 4□ 5□
	16	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including >3 observers with variable expertise	1 □ 2□ 3□ 4□ 5□
	17	Should Radiology Diagnostic Accuracy Trials report:-	
	(i)	Training and Number of Investigators?	1 □ 2□ 3□ 4□ 5□
	(ii)	Whether extra training was provided for a new technique?	1 □ 2□ 3□ 4□ 5□
	(iii)	Whether readers were blinded to prior test results?	1 □ 2□ 3□ 4□ 5□
	(iv)	Whether readers were blinded to clinical information?	1 □ 2□ 3□ 4□ 5□
RESULTS	18	Should Radiology Diagnostic Accuracy trials report study flow, including eligible patients who did not undergo Index or Reference tests, and explain why?	1 □ 2□ 3□ 4□ 5□
	19	Should a flow diagram be provided?	1 □ 2□ 3□ 4□ 5□
	20	Should changes in diagnosis after each test be reported?	1 □ 2□ 3□ 4□ 5□
	21	Should patient demographics provided including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies be provided?	1 □ 2□ 3□ 4□ 5□
	22	Should the severity (spectrum) of the disease entity be explicitly reported?	1 □ 2□ 3□ 4□ 5□
	23	Should the time difference between the index test and reference standard be provided?	1 □ 2□ 3□ 4□ 5□
	24	Do you need to know if any other treatments were provided between the two tests?	1 □ 2□ 3□ 4□ 5□
Adverse Events	25	Adverse events should be reported for either the index test or reference standard	1 □ 2□ 3□ 4□ 5□
	26	Should it be stated if contrast could have been avoided?	1 □ 2□ 3□ 4□ 5□

	27	Would you need to know if the level of radiation was less for the index test?	1 □	2□	3□	4□	5□
	28	Should Radiology Diagnostic Accuracy trials include a cross tabulation of results of index and reference tests?	1 □	2□	3□	4□	5□
STATISTICS	29	Radiology Diagnostic Accuracy trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI)	1 □	2□	3□	4□	5□
	30	Radiology Diagnostic Accuracy trials should report indeterminate/missing results and outliers; and describe how this data was handled.	1 □	2□	3□	4□	5□
	31	Radiology Diagnostic Accuracy trials should include estimates of test reproducibility	1 □	2□	3□	4□	5□
	32	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?	1 □	2□	3□	4□	5□
	(i)	Should this be provided all the time?	1 □	2□	3□	4□	5□
	33	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability	1 □	2□	3□	4□	5□
DISCUSSION	34	Should the clinical relevance of the study findings be provided?	1 □	2□	3□	4□	5□

MY RESPONSE TO CRITICAL FRIEND # 2 re: STARD_DI Delphi Round 1 (29 Sept 2014)

Please see attached STARD-DI for first round of the Delphi.

Look OK? This is to send to the REB for their approval.

Thank you both very much,

Betty Anne

STARD-DI checklist for the first round of the Delphi Technique							
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**				
Title & Abstract	1	Should Radiology Diagnostic Accuracy Trials include the words "Sensitivity & Specificity" or "Diagnostic Accuracy" in the Title/Abstract? Please choose which one below.	1 □	2□	3□	4□	5□
	(i)	"Sensitivity & Specificity" in the Title/Abstract?	1 □	2□	3□	4□	5□

	(ii)	"Diagnostic Accuracy" in the Title/Abstract?	1 □ 2□ 3□ 4□ 5□
	2	Should Radiology Diagnostic Accuracy Trials explicitly state that the Aim is to compare Index test with Reference standard for diagnosis of a specific condition	1 □ 2□ 3□ 4□ 5□
	3	What should be included in the Title and Abstract?	Please provide comment: ______________
	4	Does this need to be included in a checklist for Radiology Diagnostic Accuracy?	1 □ 2□ 3□ 4□ 5□
METHODS			
Patient selection, recruitment and sampling	5	Should Diagnostic Accuracy Trials provide Inclusion and Exclusion criteria for patients?	1 □ 2□ 3□ 4□ 5□
	6	Should Radiology Diagnostic Accuracy Trials report:-	
		Details of setting and location of study (e.g. whether primary or secondary care) ?	1 □ 2□ 3□ 4□ 5□
	(i)	Whether data collection is prospective or retrospective ?	1 □ 2□ 3□ 4□ 5□
	(ii)	Whether patient selection was consecutive or non-consecutive?	1 □ 2□ 3□ 4□ 5□
	(iii)	Sample size and limitations?	1 □ 2□ 3□ 4□ 5□
	(iv)	Whether a Power calculation was performed?	1 □ 2□ 3□ 4□ 5□
	(v)	Start and End dates of the study?	1 □ 2□ 3□ 4□ 5□
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**
	7	Should a thorough description of the nature of the disease be presented?	1 □ 2□ 3□ 4□ 5□
Diagnostic test methods	8	Should Radiology Diagnostic Accuracy Trials explicitly state the Reference Standard and its rationale?	1 □ 2□ 3□ 4□ 5□
Imperfect Reference Standard	9	If the Reference Standard is unavailable or imperfect, should trials state what alternative was used and justify the alternative? For example: New MRI technique for rotator cuff tear (index test) compared to arthroscopy (reference standard). If only 10% went on to surgery this would create a verification bias. Test only works in patients with severe disease creating a verification bias. Therefore would a mixed standard of reference be required if several orthopedic surgeons think there is no rotator cuff injury?	1 □ 2□ 3□ 4□ 5□

	10	If you do not think a mixed standard or panel standard would be of benefit, please insert your comments of what would be a viable alternative for diagnostic studies with an imperfect reference standard?	Please provide comment: _______________
Test	11	Technical specifications for index and reference tests should be reported:-	
	(i)	In All Diagnostic Accuracy Trials?	1 □ 2□ 3□ 4□ 5□
	(ii)	Only where the Test is not the standard of care?	1 □ 2□ 3□ 4□ 5□
	12	Should Radiology Diagnostic Accuracy trials explicitly state generalisability of the technique to other vendor equipment?	1 □ 2□ 3□ 4□ 5□
	13	A sound theoretical Physics basis of the test should be provided:	
	(i)	For all techniques used?	1 □ 2□ 3□ 4□ 5□
	(ii)	Only for New Techniques?	1 □ 2□ 3□ 4□ 5□
	14	If the test was modified during the study, do you need to know?	1 □ 2□ 3□ 4□ 5□
Section & Topic	**Item #**		**Likert Rating Scale: Please rate your response whereby 1- least agree to 5 strongly agree**
Analysis	15	Should Radiology Diagnostic Accuracy Trials report Cut-off values or specific diagnostic criteria for Index and Reference tests?	1 □ 2□ 3□ 4□ 5□
	16	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including >3 observers with variable expertise	1 □ 2□ 3□ 4□ 5□
	17	Should Radiology Diagnostic Accuracy Trials report:-	
	(i)	Training and Number of Investigators?	1 □ 2□ 3□ 4□ 5□
	(ii)	Whether extra training was provided for a new technique?	1 □ 2□ 3□ 4□ 5□
	(iii)	Whether readers were blinded to prior test results?	1 □ 2□ 3□ 4□ 5□
	(iv)	Whether readers were blinded to clinical information?	1 □ 2□ 3□ 4□ 5□

RESULTS	18	Should Radiology Diagnostic Accuracy trials report study flow, including eligible patients who did not undergo Index or Reference tests, and explain why?	1 □ 2□ 3□ 4□ 5□
	19	Should a flow diagram be provided?	1 □ 2□ 3□ 4□ 5□
	20	Should changes in diagnosis after each test be reported?	1 □ 2□ 3□ 4□ 5□
	21	Should patient demographics provided including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies be provided?	1 □ 2□ 3□ 4□ 5□
	22	Should the severity (spectrum) of the disease entity be explicitly reported?	1 □ 2□ 3□ 4□ 5□
	23	Should the time difference between the index test and reference standard be provided?	1 □ 2□ 3□ 4□ 5□
	24	Do you need to know if any other treatments were provided between the two tests?	1 □ 2□ 3□ 4□ 5□
Adverse Events	25	Adverse events should be reported for either the index test or reference standard	1 □ 2□ 3□ 4□ 5□
	26	Should it be stated if contrast could have been avoided?	1 □ 2□ 3□ 4□ 5□
	27	Would you need to know if the level of radiation was less for the index test?	1 □ 2□ 3□ 4□ 5□
	28	Should Radiology Diagnostic Accuracy trials include a cross tabulation of results of index and reference tests?	1 □ 2□ 3□ 4□ 5□
STATISTICS	29	Radiology Diagnostic Accuracy trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI)	1 □ 2□ 3□ 4□ 5□
	30	Radiology Diagnostic Accuracy trials should report indeterminate/missing results and outliers; and describe how this data was handled.	1 □ 2□ 3□ 4□ 5□
	31	Radiology Diagnostic Accuracy trials should include estimates of test reproducibility	1 □ 2□ 3□ 4□ 5□
	32	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?	1 □ 2□ 3□ 4□ 5□
	(i)	Should this be provided all the time?	1 □ 2□ 3□ 4□ 5□
	33	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability	1 □ 2□ 3□ 4□ 5□
DISCUSSION	34	Should the clinical relevance of the study findings be provided?	1 □ 2□ 3□ 4□ 5□

At First glance this looks fine. But I will find a wifi hotspot this weekend, have a detailed look, and get back to you.

Thanks very much and sorry for the delay

Message to Critical Friend
Fri 2014-11-07 14:28

I hope this email finds you well. I am heading to the UK Saturday night –arriving Sunday morning.

I'll be near Middlesex University at the following address:

Heatherbank Guest House
25-27 Woodside Park Rd
Finchley
Barnett, N12 8RT, United Kingdom
+442084469403

It would be great to meet for a coffee but I don't know if that's doable for you. I'm attending a ceremony on Wednesday night and I'll be leaving next Thursday.

Just to give you an update on the study - round 1 of the Delphi has gone out and 3/8 have responded thus far. There are a few away.

If meeting is not an option, I'll email you once I have more study information to share with you.

Thank you,

From Critical Friend # 2
November 23, 2014 10:28 PM

So sorry - this e mail got buried under about 150 unanswered messages & I am still about 3 weeks behind with answering.

I hope you had a good trip to the UK. The weather is certainly milder if nothing else! Norh London is great - my favourite part. I lived in Golders Green for a few years.

It turns out that I would not have been able to meet you during your stay in any case - I am back to working Resident-style 70+ hour weeks because the hospital is understaffed. Not pleasant & hence the administrative mess! But I am catching up :)

Glad to hear you have had a good initial response. It would be useful to have a telephone conversation at some point to clear up a couple of things. The time difference means this would have to be in the afternoon some time for you.

I will contact you to arrange this in the next couple of weeks as I try to get on top of things here. In the meantime please keep me updated as to your progress.

Best wishes

MY RESPONSE TO MENTOR & CRITICAL FRIEND # 2 re: Format of STARD_DI Delphi Round 1 (12 Dec 2014)

Hi Wael & Nitin

Please see the results from Delphi Round 1. Nitin, Wael and I discussed removing the responses for items that scored 100% from the panel for the next round of the Delphi. However; where the responses are at 70% panel agreement, a positive consensus can still be defined as described by Nieuwenhuijze et al., (2014). Whereas; a negative consensus was described as < 70% panel agreement.

Hence; items that scored < 70% would be added to the next round and the respondents will be asked to rank order the items they deem to be a priority. This will render patterns of agreement and disagreement as consensus of the revised tool develops (Hsu et al., 2007).

In addition, the respondents will also be provided a summary of the responses from the round 1 of the Delphi in case they wish to change their responses based on their cohort's responses (anonymized).

Does this seem acceptable to you both?

Thank you,

Appendix 25i: Response from Critical Friend # 2 re: Format of STARD_DI Delphi Round 2 (25 Jan 2015)

Just to clarify:

"It is worth considering:-
(i) Presenting supporting evidence for inclusion of each contentious item in Round 2, to ensure the panel understand the question.

Items sent back to the panel were based on the prior tool (STARD). Therefore it's not based on supporting evidence per individual item at this point of development."

By "supporting evidence" I actually meant referring to original research (i.e. published systematic reviews of sources of bias and variation in diagnostic test accuracy studies) to provide reasons/background to the panel as to why certain items might be included in the tool. All of the items in STARD and QUADAS were based on evaluation of the relevant literature, and the panel in QUADAS were given summarised evidence for each of the items included in their checklist.
http://www.bris.ac.uk/media-library/sites/quadas/migrated/documents/originalquadas-bmc.pdf
(see p3, Delphi Round 1)

However, it is fair to say that most if not all of the contentious items do not really have any supporting evidence in the literature evaluating sources of bias and variation in diagnostic accuracy studies. The only possible exception is Q32 (Should changes in diagnosis be reported after each test) which could possibly result in partial verification bias. However, for just 1 item I would tend to agree that it's not congruent to present this data as was done in QUADAS.

Best wishes

MY RESPONSE TO CRITICAL FRIEND # 2 re: Suggestions for Format of STARD_DI Delphi Round 2 (27 Jan 2015)

Thanks again for your suggestions and expert advice. With respect to supplying supporting evidence to the panel, Whiting et al. did this based on the results of their systematic reviews.

My doctoral study is being conducted in follow-up to my own systematic review (master's thesis) which examined adherence for both the STARD and the QUADAS for abdominal imaging trials in the ED setting pre + post the development of both tools.

I found improvement for some items for both tools but I also noted that some items were not pertinent to radiology arguing that a 'one size fits all' doesn't always work and that a tool more specific to radiology diagnostic accuracy may render further benefit.

Certainly, once this new tool is developed and I will write up the explanatory document to be used with the new tool with the aim of facilitating its use, understanding and dissemination of the new tool/checklist. This document will provide clarification of the meaning, rationale and optimal use for each item on the checklist. As well I will provide a short summary of the available evidence on bias and applicability for each item as described by Bossuyt at el. (2003).

http://www.clinchem.org/content/49/1/7.full.pdf+html

Thank you,

Betty Anne

Message to CRITICAL FRIEND #2
Wed, 1 Apr 2015 15:05:15 +0000

Dear Nitin,

RE: 20140142-01H
"Developing a Standardized Tool for Interpretation of Radiology Diagnostic Accuracy Trials"
I would like to sincerely thank you for your time and expertise collaborating with me as my critical friend on my doctoral study.

Please find below the revised tool that has been developed in conjunction with the radiological experts who provided their anonymized responses when completing the Delphi technique. Consensus was defined as meeting 70% agreement on an individual item.

As this revised STARD was developed so that it is specific to radiology I have named the tool the "**RADART**" - **R**adiology **D**iagnostic **A**ccuracy **R**eporting **T**ool.
Based on the rounds from the Delphi there were 13 items from the STARD_DI that did not meet consensus <70% agreement. Notably 2 of the items from the current STARD are not in the RADART.

Methods for calculating test reproducibility and the provision of a cross tabulation of the results of the index tests and reference standard.

RADART –Radiology Diagnostic Accuracy Reporting Tool

Item	
Title & Abstract 1.	**Radiology Diagnostic Accuracy Trials should include the words "Diagnostic Accuracy" in the Title/Abstract.**
2.	**Radiology Diagnostic Accuracy Trials should explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition**.
3. METHODS Patient selection, recruitment & sampling	Radiology Diagnostic Accuracy Trials should provide inclusion and exclusion criteria. Details of setting and location of study (e.g. **whether primary or secondary care**) should be provided.
4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.**
5.	**Data collection (prospective or retrospective) should be provided with start dates and end dates provided.**
6.	Whether patient selection was consecutive or non-consecutive?
7.	**Sample size and limitations should be provided.**
8. Diagnostic test methods	Radiology Diagnostic Accuracy Trials should explicitly state the reference standard and its rationale.

9. **Imperfect reference standard**	**If the Reference Standard is unavailable or imperfect, use of an alternative reference standard should be should justified.**
10. **Test**	**Technical specifications for the index test and reference test should be reported in all radiology diagnostic accuracy studies.**
11.	**Sound theoretical physics basis of the index test should be provided for new techniques.**
12.	**Modifications during the study should be reported if they occurred.**
13. **Analysis**	Radiology Diagnostic Accuracy Trials should report cut-off values **for specific diagnostic criteria** for index and reference tests.
14.	The training and number of investigators should be described including **any extra training for new techniques.**
15.	Whether readers were blinded to prior test results or clinical information should be known.
16. **RESULTS**	Radiology Diagnostic Accuracy Trials report study flow with a flow diagram, including eligible patients who did not undergo index or reference tests, and provide explanations.
17.	Patient demographics including age, sex, presenting diagnosis or symptoms, any co-morbidities, and concurrent therapies should be provided.
18.	The severity (spectrum) of the disease entity should be explicitly reported.
19.	The time difference between the index test and reference test should be provided.
20.	Details of any other treatments provided between the two tests should be described.
21. **Adverse Events**	Adverse events should be reported for either the index test or reference standard.
22. **STATISTICS**	Radiology Diagnostic Accuracy Trials one should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI).
23.	Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers; and describe how this data was handled.
Item	
24.	Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability.

25.	Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise.
26.	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.
27. **DISCUSSION**	The clinical relevance of the study findings should be provided.

Items in bold (1, 2, 3 (partial), 4, 5, 7, 9, 10, 11, 12 and 13 (partial) and 14 (partial) are specific to the RADART Reporting Tool.
Bossuyt (2003)
All results will be elaborated and explained in great detail as I continue with the write up.

I am now writing the explanatory document which will aid the reader in facilitating the use of this checklist plus it will aid comprehension of the results as reported within the article (radiology diagnostic accuracy study).
Once the Amendment for the new tool (RADART) has been approved by the REB I will be testing it with the residents and Fellows. I will share those results with you at a later date.
Thank you,
Betty Anne

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W. For the STARD Group (2003) 'Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative', *Radiology*, 226 (1), pp. 24-28.

Appendix 25j: Response from Critical Friend # 2 re: RADART (26 April 2015)

Hi Betty Anne

Do you have the results of Delphi Round 2? Could you possibly forward it to me?
I don't think I ever saw them and I am just trying to follow through exactly how the data went through after round one.

I think the provisional RADART below looks good, although I do not understand #4 and I do have some suggestions and comments.
However, before sending you a full response and the excel sheet I have written, I would like to see the Round 2 results if available.

MY RESPONSE TO CRITICAL FRIEND # 2 re: RADART (27 Apr 2015)
MESSAGE SENT TO CRITICAL FRIEND #2

Hi Nitin,

Thanks very much for your email.

The results of Delphi Round 1 were sent out to the panel offering them the opportunity to change/alter their original response based on those of their cohort. The number of items that scored greater than 70% positive rate did not change from the first round of the Delphi 1 however; the preponderance of positivity between agree or strongly agree changed for a few items as listed below. In the revised responses for the Delphi 1 it states that there were 9 responders. It should be clarified that only member of the expert panel chose to change their when accessing the same survey resulting in a total of nine responders. We were not able to identify the change or duplication due to the anonymity of the survey.

5.5.3 STARD Delphi 1 Revised Reponses – Consensus Not Met

The following questions were not agreed upon by the responders from the first round of the STARD_DI Delphi. The same Likert Scale was provided: strongly disagree, disagree, neither agree, nor disagree, agree, strongly or agree. These results indicated the percentage of responses that were not in agreement by the cohort with a score of <70% for agree and strongly agree.
STARD_DI Delphi Round 1 Results: *Responders: N = 8 Questions skipped: N = 0*

STARD_DI Delphi Round 1 Revised Reponses: *Responders: N = 9. Questions skipped: N = 1*

Item		Response N = 8	Response N = 9
Title & Abstract	"Sensitivity and Specificity" in the Title/Abstract?	50%	56%
Patient Selection, recruitment & sampling	Whether a power calculation was performed?	38%	44%
	Should a thorough description of the nature of the disease be presented?	38%	44%
	Only when the test is not the standard of care?	38%	33%
	Should Radiology Diagnostic Accuracy Trials explicitly state generalisability of the technique to other vendor equipment?	63%	67%
	For all techniques used?	25%	33%
Analysis	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including > 3 observers with variable expertise?	50%	44%
RESULTS	Should changes in diagnosis after each test be reported?	50%	56%
	Should it be reported if contrast could have been avoided?	50%	56%
	Would you need to know if the level of radiation was less for the index test?	50%	67%
	Should Radiology Diagnostic Accuracy Trials include a cross tabulation of results of index and reference tests?	63%	67%
STATISTICS	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.	63%	67%
	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?	25%	22%
	Should this be provided all the time?	13%	11%
	Technical specifications for the index test and reference tests should be reported only when the test is not standard of care?	38%	33%

(Bossuyt et al., 2003)

5.5.4 STARD_DI Additional Items in the Title and Abstract

When asked “what should be included in the Title and Abstract? The one additional response was:

1. Clinical impact, adaptability to different population

5.5.5 STARD_DI Additional Items Recommended by the Expert Panel

The following response rate is in response to the qualitative response listed above with the results from the STARD_DI Delphi 1 compared to the revised response when the survey was completed again.

STARD_DI Delphi Round 1 Results: *Responders: N = 8 Questions skipped: N = 0*

STARD_DI Delphi Round 1 Revised Reponses: *Responders: N = 9. Questions skipped: N = 1*

Item		Response N = 8	Response N = 9
Title	Does this need to be included in a checklist for Radiology Diagnostic Accuracy?	63%	44%

5.6 STARD_DI Delphi Round 2

Based on their responses from the Delphi Round 1 there were seven questions in the second round of the Delphi whereby the participants were asked to choose three out of the seven items they thought should remain in the revised checklist for the STARD_DI.

In addition they were also provided the opportunity to add comments stating their rationale concerning the items they deemed to rate as a priority amongst the items reviewed.

Only the items that did not reach consensus were provided to the participants via Survey method. The following two items were agreed upon as the combined responses for agree and strongly agree were > 70%. The same Likert Scale was provided as previously described whereby the radiological expert rated each item according to the rating scale of strongly agree, disagree, neither agree, nor disagree, agree, strongly or agree.
The survey responses for the STARD_DI Delphi Round 2 Revised can be found in Appendix 3.

5.6.1 STARD_DI Delphi Round 2 Consensus Met: *Responders: N = 8 Questions skipped: N = 0*

Item		**Response N = 8**
	Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise	75%
Rationale	1. Inter-observer agreement is valuable, but as long as the level of expertise is explicitly stated. It should not be mandatory to provide more than two observers. 2. Ideal but sometimes not practical.	
	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility	83.3%
Rationale	1. None provided	

5.6 STARD_DI Delphi Round 2

Based on their responses from the Delphi Round 1 there were seven questions in the second round of the Delphi whereby the participants were asked to choose three out of the seven items they thought should remain in the revised checklist for the STARD_DI.

In addition they were also provided the opportunity to add comments stating their rationale concerning the items they deemed to rate as a priority amongst the items reviewed.

Only the items that did not reach consensus were provided to the participants via Survey method. The following two items were agreed upon as the combined responses for agree and strongly agree were > 70%. The same Likert Scale was provided as previously described whereby the radiological expert rated each item according to the rating scale of strongly agree, disagree, neither agree, nor disagree, agree, strongly or agree.

The survey responses for the STARD_DI Delphi Round 2 Revised can be found in Appendix 3.

5.6.1 STARD_DI Delphi Round 2 Consensus Met: *Responders: N = 8 Questions skipped: N = 0*

Item		**Response N = 8**
	Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise	75%
Rationale	3. Inter-observer agreement is valuable, but as long as the level of expertise is explicitly stated. It should not be mandatory to provide more than two observers. 4. Ideal but sometimes not practical.	
	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility	83.3%
Rationale	2. None provided	

5.6.1 STARD_DI Delphi Round 2 Consensus Not Met: *Responders: N = 8 Questions skipped: N = 0*
The remaining five items did not reach consensus as the response rate was <70%.
STARD_DI Delphi Round 2 Results: *Answered: N = 6 Questions skipped: N = 2*

Item	**Response N = 6**

	Whether a power calculation was performed	50%
Comment	1. Need to know how many patients required to have statistical significance. 2. Not always possible to provide a power calculation, especially for new diagnostic techniques or for populations who have not been evaluated previously. 3. Statistical game sometimes.	
	Radiology Diagnostic Accuracy Trials should explicitly state generalisabilty of the technique to other vendor equipment for all techniques used.	20%
Comment	1. A study shouldn't be done unless the test can be generalized to all vendor platforms. 2. Readers should be alerted if the performance of the technique is vendor specific or limited to a specific population. 3. May be difficult to guess sometimes.	37.5%
	Changes in the diagnosis after each test should be reported.	60%
Comment	1. There were no responses.	
	You should know if the level of radiation was less for the index test.	33.3%
Comment	1. Should be reported if test involves radiation. 2. Usually not a game changer.	
	Radiology Diagnostic Accuracy Trials should explicitly state generalisabilty of the technique to other vendor equipment for all techniques used.	66.7%
Comment	1. I find this question confusing.	

Thank you,

Betty Anne

Appendix 25k: Response from Critical Friend # 2 re: RADART (28 Apr and 7 May 2015)

Thanks very much for this. I will get back to you later in the week.

FOLLOW-UP MESSAGE FROM CRITICAL FRIEND #2
Thu 2015-05-07 21:25

Thanks very much for your response.
I am impressed with your results and the phenomenal timeframe in which you are achieving your goals. Very well done.

Hi Betty Anne

I have followed through your method in some detail (hence the delayed response as this is quite time-consuming). I am pleased that you went on to a Round 2 and used "Include/Exclude" rather than Likert scales; this was exactly right.

On the basis of my reading of your study, I do have a few points to make.

1. Original Round 1: Lack of inclusion of "Patient recruitment method" (this was present in STARD)
- You can suggest that this was not really applicable to radiology. This is not really a big deal - I just noted that it wasn't represented on your original questionnaire and was included in the original STARD.

2. Revised Round 1
(i) **Stability of Responses.** This is a key ingredient in achieving consensus and you did demonstrate stability for all of the items which did not reach consensus except 2 items (Radiation exposure being the most relevant). Technically this is unstable and consensus cannot be reliably assessed according to some authors. However, you did include this in Round 2 and it was not selected by the panel, so I think this is OK.
(ii) **Lack of Consensus.** I counted **12 items** which did not meet consensus. This agrees with your information. But only 6 items were included in Round 2 ?? (See next section)

3. Round 2

(i) You included 6 items in Round 2 (out of the 12 items not reaching consensus in Round 1). **Why were 6 non-consensus items excluded and 6 included?** As you know, this is a potential source of bias unless there was a specific reason for excluding those 6 items. See my spreadsheet, columns U-W.

(ii) You asked the panel to **arbitrarily select 3 of the 6 for inclusion**, rather than simply score them all as "Include/Exclude". What was the basis for this? Why not assess what consensus was reached for all 6 items individually and exclude those not reaching consensus (as for previous rounds) ? To cut down the list, you could then have done a Round 3 asking respondents to include the 3 most important.

Again, I don't think this is a particularly big deal. I assume the reason is space constraints in the RADART, but I would be grateful if you could clarify.

4. Final RADART

(i) Item #4**: Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.**

I am not sure where this has come from? This question was not included in any of the Delphi Rounds (e.g. your document STARD_DI Delphi Round 1. pdf).

It is also not a particularly useful item - by definition for a Diagnostic accuracy trial all participants should have received both?

(ii) You have identified several strands in RADART which are new and specific, and different or modified from STARD. But my reading of the differences is not the same as yours.

As far as I can see (please refer to my spreadsheet showing the step by step evolution of RADART from STARD) there are **5 items which are new to RADART**:-

- Sample size and limitations
- Justify alternative reference standards used if original was unavailable
- Theoretical basis for new techniques
- Modifications to tests during study
- extra training for new tests

.... and **3 items which are modified from STARD to be more specific to Radiology**

- Include "Diagnostic Accuracy" in title
- Explicitly state aim to compare index and ref test
- Estimate interobserver reliability & robust assx with >3 observers

This is just my reading of your study.

My overall impression is that you have done a fantastic job in a short time with this. The queries and critical points above are just my opinion.

The only thing I am concerned about is why only some of the non-consensus items from Revised Round 1 were included in Round 2. The rest are minor points.

I do think it is excellent that most of your responses demonstrated stability and shows good validity.

RADART looks like it could be a really useful tool - the validated inclusion of inter-observer reliability estimates is particularly interesting and clearly signals a remodelling and re-emphasis from the generic STARD criteria to be more specific to Radiology.

Very well done with this.

MY RESPONSE TO CRITICAL FRIEND # 2 re: DEVELOPMENT OF RADART

Dear Nitin,

Thank you very much for your thoughts and support throughout the development of this new tool.

With respect to **"Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard",** this is explained in the elaboration document.

4.	**Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.**

With respect to trial recruitment the STARD statement recommends providing information on recruitment of patients that is based on presenting symptoms, previous investigations or results obtained from either the index test or reference standard (Bossuyt et al., 2003b) Ultimately, the provision of considerable detail is required to determine if the patients included in the trial not only met eligibility criteria but it is also relevant to know if they were appropriately chosen for the index test under study (Riegelman, 2000).

Similar to the STARD the RADART recommends trial recruitment should include both presenting symptoms and results from previous tests however; these items ranked lower priority than whether the participant received either the index test or reference standard. For example, a study comparing the diagnostic accuracy of metal artifact reduction sequence magnetic resonance imaging (MARS MRI) and ultrasound to detect painful metal on metal arthroplasty was recently published. For this particular diagnostic accuracy study participant recruitment included patients meeting the inclusion criterion as well as having undergone a MARS MRI within the previous year and who then subsequently went on to be protocolled to have an (ultrasound scanning) USS (index test). Whereas; newer patients who presented with pain were also included to receive both the prospective MARS MRI and USS (Siddiqui et al., 2014).

The assigning of participants into a diagnostic accuracy study can occur in one of the three following ways:

1. Participants who fulfill the eligibility criterion and target population of study are identified prior to undergoing the index test and reference standard.
2. Participants with and without disease are identified based on the results of their reference standard and who then undergo the index test.
3. Participants who have undergone the index test are identified to then undergo the reference standard.

In these three different scenarios, option one is considered the best method for patient recruitment as those identified represent the target population. Alternatively, option two and three may also be used however; each can result in biases from either of these methods. In option two if the investigator chooses only those who meet a clear diagnosis of the disease rather than including those in the grey area such as those with concomitant co-morbidities the results may result in spectrum bias due to the exclusion of those with co-morbidities and may have also have had a positive index test. Whereas; option three may also result in bias as the identification of study participants was based on findings from the index test which can result in verification bias (Riegelman, 2013).

In an effort to reduce the incidence of bias the RADART recommends including participants based on those meeting eligibility criterion of whether or not the patient had undergone either the index test or reference standard. To clarify; participants in diagnostic accuracy studies undergo testing with both the index test and reference standard to compare the results. However; for this particular item of the RADART it is recommending to include participants based on whether or not they had received the index test or reference standard pending the medical condition under study.

The Delphi Rounds are done now so I can't go back.

Plasticity of responses - please see attached Chapter Wael had sent me.

Thanks again,

Betty Anne

Dear Nitin,

Thank you once again for your very detailed assessment of my work in creating the RADART with the radiological experts.

I am currently finishing the elaboration document which describes each item of the RADART with mention of the similarities to the STARD where applicable.

I am including radiological examples for each item making the list specific to radiology. Then there are some items that are new to the RADART and hence stand alone.

Please find my responses embedded below to your queries:

Dear Betty Anne

Thanks very much for your response.
I am impressed with your results and the phenomenal timeframe in which you are achieving your goals. Very well done.

Many thanks to you as well.

I have followed through your method in some detail (hence the delayed response as this is quite time-consuming). I am pleased that you went on to a Round 2 and used "Include/Exclude" rather than Likert scales; this was exactly right.

Many thanks again.

On the basis of my reading of your study, I do have a few points to make.

1. Original Round 1: Lack of inclusion of "Patient recruitment method" (this was present in STARD; see row 14 on my spreadsheet in green)
- You can suggest that this was not really applicable to radiology. This is not really a big deal - I just noted that it wasn't represented on your original questionnaire and was included in the original STARD.

Although most items of the Needs Assessment were met with consensus by the cohort, there were several items that were not.

Q3. Recruitment of participants into the trial. Please rate 1-3 which items you deem most essential?

presenting symptoms: 1.88 (average ranking)
results from previous tests:1.38 (average ranking)
participants had received either the index test or the reference standard 2.75 (average ranking).

Hence whether participants had received either the index test or the reference standard was maintained as two remaining items were deemed not as applicable. I see your point though in saying index test or reference standard which seems counterintuitive as participants need to receive both in a DA study. This will be iterated in the elaboration document.

2. Revised Round 1
(i) **Stability of Responses.** This is a key ingredient in achieving consensus and you did demonstrate stability for all of the items which did not reach consensus except 2 items (Radiation exposure being the most relevant). Technically this is unstable and consensus cannot be reliably assessed according to some authors. However, you did include this in Round 2 and it was not selected by the panel, so I think this is OK.
(ii) **Lack of Consensus.** I counted **12 items** which did not meet consensus. This agrees with your information. But only 6 items were included in Round 2 ?? (See next section)

3. Round 2
(i) You included 6 items in Round 2 (out of the 12 items not reaching consensus in Round 1). **Why were 6 non-consensus items excluded and 6 included?** As you know, this is a potential source of bias unless there was a specific reason for excluding those 6 items. See my spreadsheet, columns U-W.
(ii) You asked the panel to **arbitrarily select 3 of the 6 for inclusion**, rather than simply score them all as "Include/Exclude". What was the basis for this? Why not assess what consensus was reached for all 6 items individually and exclude those not reaching consensus (as for previous rounds) ? To cut down the list, you could then have done a Round 3 asking respondents to include the 3 most important.
Again, I don't think this is a particularly big deal. I assume the reason is space constraints in the RADART, but I would be grateful if you could clarify.

An excellent question and thank you for bringing this up as I need to elaborate on this in the Chapter. The decision on which items to choose to include was decided in conjunction with Dr. Shabana.

We carefully examined all items that did not reach consensus from Delphi Round 1 (<70%) and chose 7 of the items for Round 2 of the Delphi. The respondents were asked to pick their top three as we did not want them to include all of the items to mitigate a plasticity in response which can sometime occur with survey responders.

STARD_DI Delphi Round 1 Revised Reponses: *Responders: N = 9. Questions skipped: N = 1*

Item		Response N = 8	Response N = 9
Title & Abstract	"Sensitivity and Specificity" in the Title/Abstract? not included as we already knew from Round 1 that the respondents favored the use of "Diagnostic Accuracy" in the title which is actually huge as retrieving diagnostic articles using the search "sensitivity and specificity" as a Mesh heading is criticized in the literature as many articles are missed when one performs a literature search for DA studies.	50%	56%
Patient Selection, recruitment & sampling	Whether a power calculation was performed? Included in Round 2.	38%	44%
	Should a thorough description of the nature of the disease be presented? Not included - deemed not applicable as respondents already indicated in Round 1 that the purpose of a DA study is to compare the index test with a reference standard for diagnosis of a specific condition. In addition the condition itself would most likely be described in the introduction of a DA study.	38%	44%
	Only when the test is not the standard of care? This is in reference to the need for describing the techical requirement for the index test which is agreed upon. Therefore this question was not included as we already had consensus on this key requirement for radiology diagnostic accuracy studies.	38%	33%
	Should Radiology Diagnostic Accuracy Trials explicitly state generalisability of the technique to other vendor equipment? Included	63%	67%
	For all techniques used? Not included - as it relates to the above question which is already being asked.	25%	33%
Analysis	Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including > 3 observers with variable expertise? Included	50%	44%
RESULTS	Should changes in diagnosis after each test be reported? Included	50%	56%
	Should it be reported if contrast could have been avoided? Not included - Dr. Shabana "not applicable"	50%	56%
	Would you need to know if the level of radiation was less for the index test? Included	50%	67%

	Should Radiology Diagnostic Accuracy Trials include a cross tabulation of results of index and reference tests? Included	63%	67%
STATISTICS	Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility. Included	63%	67%
	To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated? Not included as we were already asking if radiology diagnostic accuracy studies should include estimates of test reproducibility.	25%	22%
	Should this be provided all the time? A dangler question on it's own- not included as we're already addressing this with the reproducbilty question.	13%	11%
	Technical specifications for the index test and reference tests should be reported only when the test is not standard of care? Not included as we knew there was aleady consenus met for including a thorough description for technical specifications for both the index test and reference standard.	38%	33%

This left us with 7 questions to include in Round 2 of the Delphi.

4. Final RADART

(i) Item #4**: Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.**
I am not sure where this has come from? This question was not included in any of the Delphi Rounds (e.g. your document STARD_DI Delphi Round 1. pdf).
It is also not a particularly useful item - by definition for a Diagnostic accuracy trial all participants should have received both?

Yes - this was worded as per the STARD item # 4

Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact the participants had received the index tests or the reference standard. I will need to clarify this in my elaboration document,

(ii) You have identified several strands in RADART which are new and specific, and different or modified from STARD. But my reading of the differences is not the same as yours.

Different or modified means an item of the RADART is similar to a particular STARD but partially different.

Example: STARD Item # 3: The study population: The inclusion and exclusion criteria, setting and locations where data were collected.

RADART: Added that whether primary or secondary care should be provided.
As far as I can see (please refer to my spreadsheet showing the step by step evolution of RADART from STARD) there are **5 items which are new to RADART**:-
- Sample size and limitations
- Justify alternative reference stds used if original was unavailable
- Theoretical basis for new techniques
- Modifications to tests during study
- extra training for new tests

.... and **3 items which are modified from STARD to be more specific to Radiology**
- Include "Diagnostic Accuracy" in title
- Explicitly state aim to compare index and ref test

- Estimate interobserver reliability & robust assx with >3 observers

This is just my reading of your study. Again you can see how I have arrived at this if you follow through my spreadsheet and look at Columns Y to AB.

Still reviewing.

My overall impression is that you have done a fantastic job in a short time with this. The queries and critical points above are just my opinion.
The only thing I am concerned about is why only some of the non-consensus items from Revised Round 1 were included in Round 2. The rest are minor points.
I do think it is excellent that most of your responses demonstrated stability and shows good validity.
RADART looks like it could be a really useful tool - the validated inclusion of inter-observer reliability estimates is particularly interesting and clearly signals a remodelling and re-emphasis from the generic STARD criteria to be more specific to Radiology.

Very well done with this.

Best wishes
Nitin

Thank you once again. I am most appreciative of your time in carefully reviewing these results and I immensely appreciate your questions and suggestions. I will keep you apprised of our next steps.

Sincerely,

Betty Anne

Appendix 25l: Response from Critical Friend # 2 re: RADART TOOL (25 May 2015)

Thanks for your response.
I can see a huge amount of thought and effort has gone into your methodology.

My points #1 and #4(i) are essentially the same. As you mention, this will need to be elaborated.
It may be worth rewording
"Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard."
to focus more specifically
"Radiology diagnostic accuracy trials should indicate whether recruitment was based on participants receiving either the index test or the reference standard"

Round 2
I think I would personally have included at least 9 (rather than 7) criteria in Round 2 - only omitting those which were autologous (as you describe).
But so long as you have an acceptable justification for the method used, there is no problem.
I think your clarifications and reasoning are fine here.

Final RADART
Please feel free to review. STARD contains some "hidden" criteria which are generic but nevertheless included in the original questionnaire.
This doesn't make any difference to RADART. I was just pointing out that some items are refinements of existing parameters rather than new criteria.

The only thing I wasn't familiar with was the concept of Plasticity of responses?
I take it this means that people tend to "spread" responses over a wide range and change between iterations, rather than congregate towards consensus?
If so, then asking respondents to choose 3/6 items seems appropriate.

Very well done again.
Best wishes

MY RESPONSE TO CRITICAL FRIEND # 2 re: SUGGESTIONS FOR RADART TOOL (12 Jun 2015)
Dear Nitin,

Thank you very much for your thoughts and support throughout the development of this new tool.

With respect to **"Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard",** this is explained in the elaboration document.

4. Radiology Diagnostic Accuracy Trials should indicate if participants had received either the index test or the reference standard.

With respect to trial recruitment the STARD statement recommends providing information on recruitment of patients that is based on presenting symptoms, previous investigations or results obtained from either the index test or reference standard **(**Bossuyt et al., 2003b). Ultimately, the provision of considerable detail is required to determine if the patients included in the trial not only met eligibility criteria but it is also relevant to know if they were appropriately chosen for the index test under study (Riegelman, 2013).

Similar to the STARD the RADART recommends trial recruitment should include both presenting symptoms and results from previous tests however; these items ranked lower priority than whether the participant received either the index test or reference standard. For example, a study comparing the diagnostic accuracy of metal artifact reduction sequence magnetic resonance imaging (MARS MRI) and ultrasound to detect painful metal on metal arthroplasty was recently published. For this particular diagnostic accuracy study participant recruitment included patients meeting the inclusion criterion as well as having undergone a MARS MRI within the previous year and who then subsequently went on to be protocolled to have an (ultrasound scanning) USS (index test). Whereas; newer patients who presented with pain were also included to receive both the prospective MARS MRI and USS (Siddiqui et al., 2014).

The assigning of participants into a diagnostic accuracy study can occur in one of the three following ways:

1. Participants who fulfill the eligibility criterion and target population of study are identified prior to undergoing the index test and reference standard.
2. Participants with and without disease are identified based on the results of their reference standard and who then undergo the index test.
3. Participants who have undergone the index test are identified to then undergo the reference standard.

In these three different scenarios, option one is considered the best method for patient recruitment as those identified represent the target population. Alternatively, option two and three may also be used however; each can result in biases from either of these methods. In option two if the investigator chooses only those who meet a clear diagnosis of the disease rather than including those in the gray area such as those with concomitant co-morbidities the results may result in spectrum bias due to the exclusion of those with co-morbidities and may have also have had a positive index test. Whereas; option three may also result in bias as the identification of study participants was based on findings from the index test which can result in verification bias (Riegelman, 2000).

In an effort to reduce the incidence of bias the RADART recommends including participants based on those meeting eligibility criterion of whether or not the patient had undergone either the index test or reference standard. To clarify; participants in diagnostic accuracy studies undergo testing with both the index test and reference standard to compare the results. However; for this particular item of the RADART it is recommending to include participants based on whether or not they had received the index test or reference standard pending the medical condition under study.

The Delphi Rounds are done now so I can't go back.

Plasticity of responses - please see attached Chapter Wael had sent me.

Thanks again,

Betty Anne

Appendix 25m: Response from Critical Friend # 2 re: METHODOLOGY (28 June 2015)

Dear Betty Anne

Thank you for your kind response and for sending the attachment on Psychology of Survey Respondents. I found this absolutely fascinating.

1. I had enquired why the Panel were asked to select 3 out of 7 items rather than score all items in Round 2. You suggested that asking the Panel to select 3 items only would reduce survey plasticity of response.
The chapter suggests that response order and context (rather than number of items) determine plasticity. So asking respondents to select 3 items would not necessarily reduce this.

However, it remains a reasonable methodology; I just think that it entails a very small chance that there might be a "hidden" 4th item not included due to the methodology.

2. Your elaboration document is very well written and nicely explains away the point that I raised. Ultimately, the Panel elected that the most important item of information to be included is whether recruitment was based on the patient having undergone the index or reference test. Whether this will "reduce the incidence of bias" (as you stated) is a slightly different point.
I am not sure it will actually "reduce" bias in my opinion - however, it allows the reader to quantify the degree of bias and is therefore a vital piece of information.

I would be happy to review your elaboration document once you have completed it.

Well done again - the excerpt from your elaboration document reads very well indeed so I am sure your thesis will be excellent.

Best wishes
Nitin

MY RESPONSE TO CRITICAL FRIEND # 2 re: Change of name from RADART to RadSTARD (29 Jun 2015)

Dear Nitin,

Thanks so much for your response and further suggestions.

I have just made a change to the RADART and changed it to the RadSTARD.

After reviewing the literature for the third time I came across a few new tools that were developed as extensions to the STARD. I do consider this tool an extension - however I do struggle with my decision on the name of this newly revised tool for although there are STARD items that the cohort did not keep - new items were added as well.

Nonetheless, please find attached my Explanation and Elaboration document.

I've changed it to the RadSTARD.

Please let me know your thoughts.

Thank you again,

Betty Anne

Appendix 25n: Response from Critical Friend # 2 re: Elaboration Document (30 June 2015)

Thanks Betty Anne

I will have a look at this over the weekend and get back to you.

Nitin

MY MESSAGE TO CRITICAL FRIEND #2 (30 Jun 2015)

Hi Nitin,

This document is currently with the REB pending approval to test with the residents and Fellows. Wael looked it over as did my Advisor from Middlesex University.

Thank you,

Betty Anne

Appendix 25o: Response from Critical Friend # 2 re: Elaboration Document (30 June 2015)

Great.
Sorry much of my responses end up being retrospective. This is due to the demands of a very heavy Fellowship at the moment.
It's still useful to get another external opinion though; it can at least provide another perspective when it comes to your write-up.
I will let you know what I think in the next few days.
Well done again on moving things through so quickly.
Best wishes
Nitin

MY MESSAGE TO CRITICAL FRIEND #2 (30 Jun 2015)

Dear Nitin,

Truthfully Wael did not me to bother you with the Elaboration Document. It took me over a month to write it.
I am proud of it. Thank you for asking to review it.
Kindest Regards,

Betty Anne

Appendix 25p: Response from Critical Friend # 2 re: Elaboration Document (17 July 2015)

Hi Betty Anne

I'm still going on this - impressive so far. It must have been a vast amount of work to write, and takes quite some time to review as well!
I will hopefully be able to finish later this week, provided my schedule isn't completely crazy.

Best wishes
Nitin

MY MESSAGE SENT TO CRITICAL FRIEND #2 (13 Jul 2015)

Hi Nitin,

Indeed it took me over a month to write it. I'm glad you’re impressed so far.
Thank you,

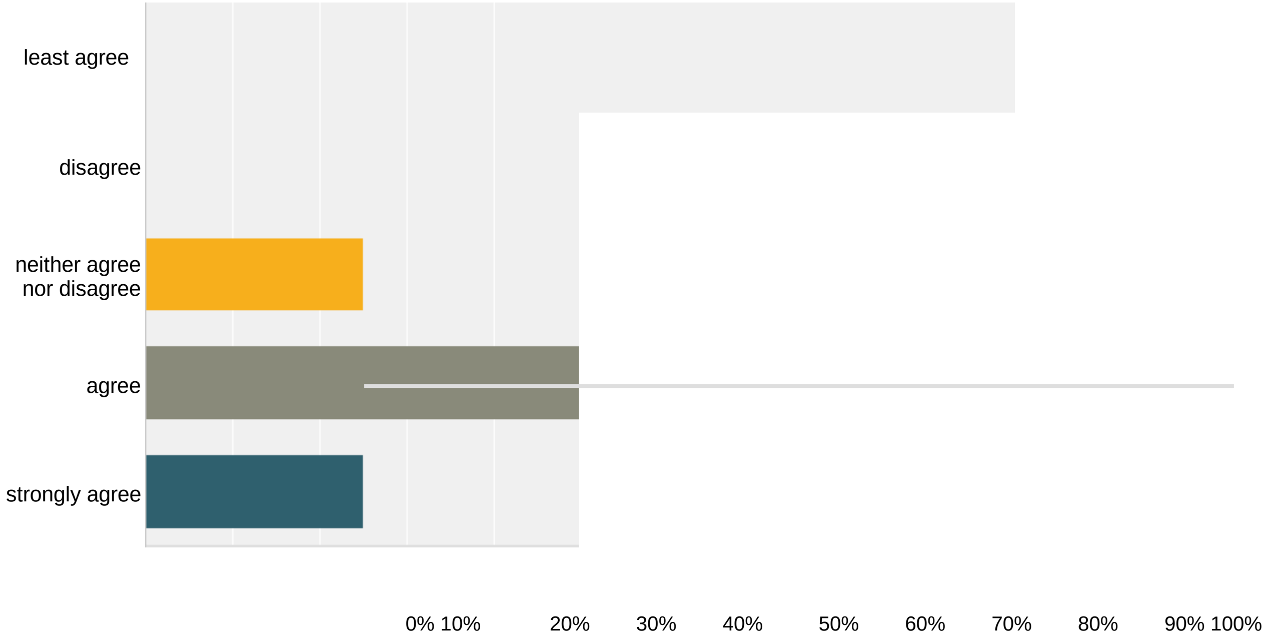
Betty Anne

Appendix 27: Results of Delphi Round 1

STARD_DI Delphi Round 1

Q1 Should Radiology Diagnostic Accuracy Trials include the words "Sensitivity & Specificity" or "Diagnostic Accuracy" in the Title/Abstract? Please choose which one below. Please rate your response whereby 1- least agree to 5 strongly agree

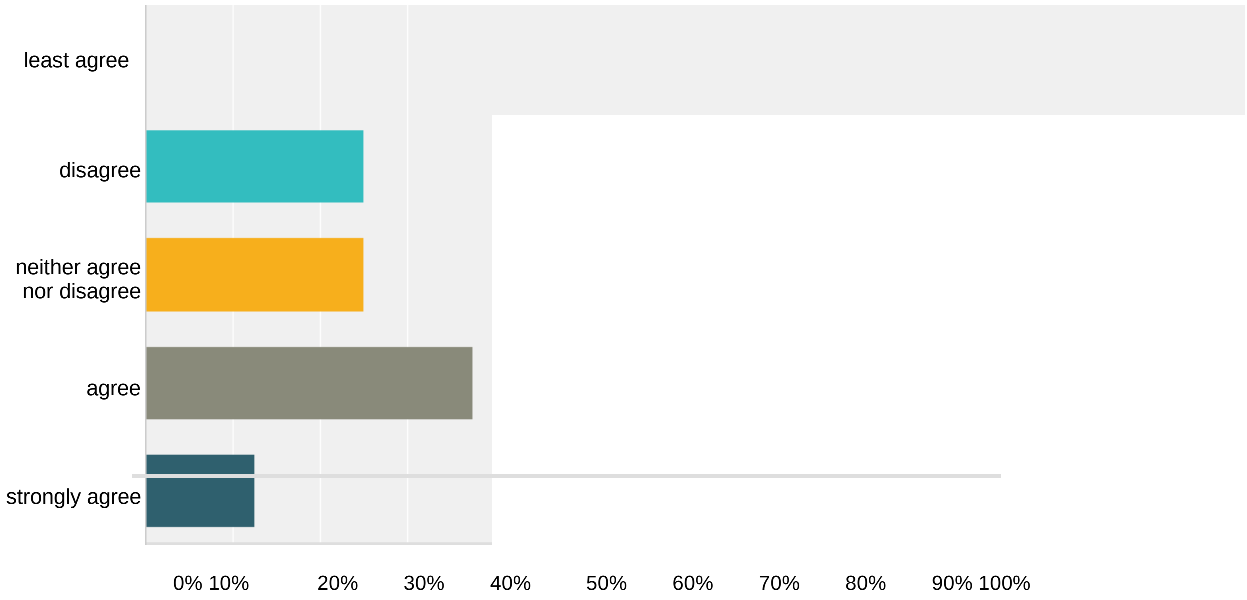
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	50.00%	4
strongly agree	25.00%	2
Total Respondents: 8		

Q2 "Sensitivity & Specificity" in the Title/Abstract?

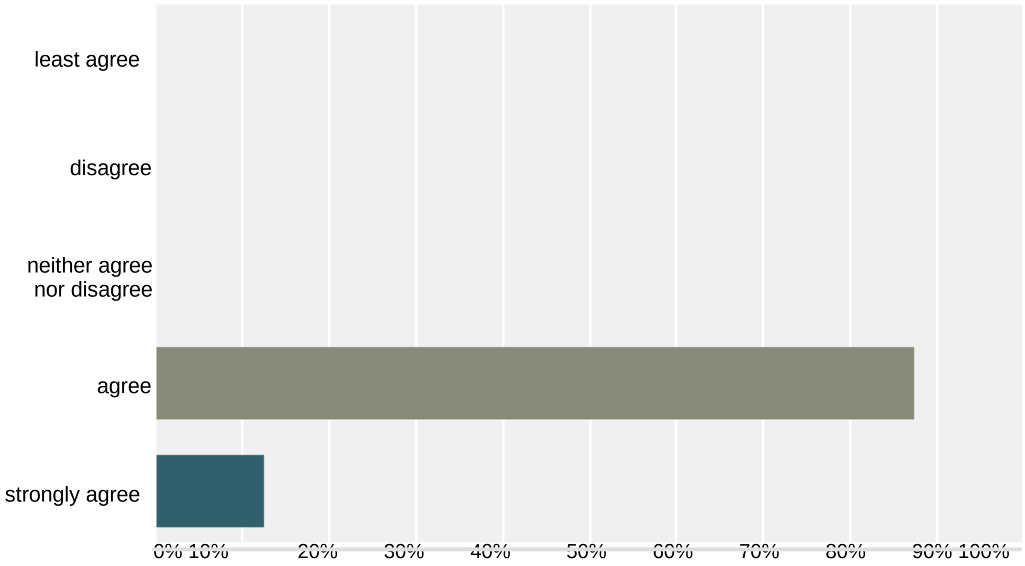
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	25.00%	2
neither agree nor disagree	25.00%	2
agree	37.50%	3
strongly agree	12.50%	1
Total Respondents: 8		

Q3 "Diagnostic Accuracy" in the Title/Abstract?

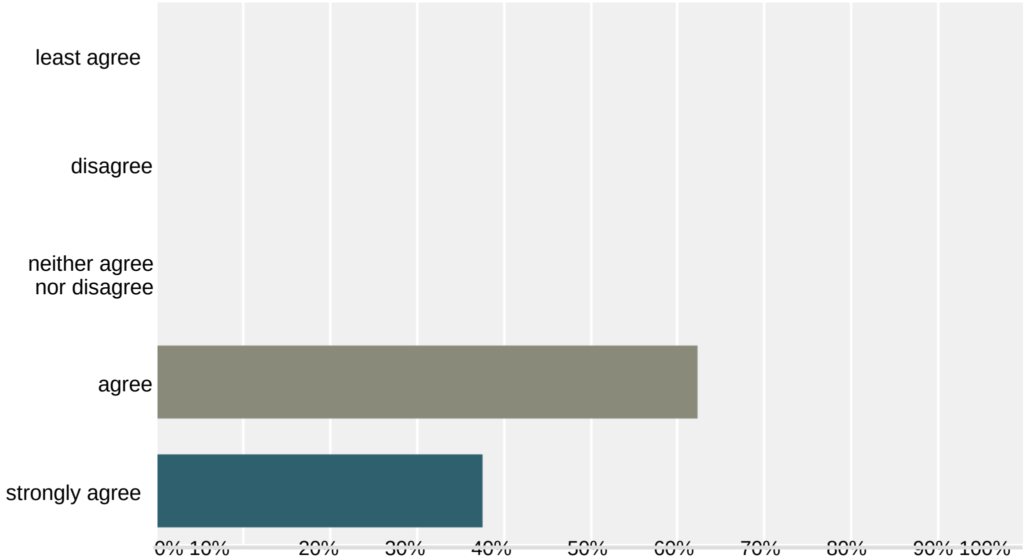
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	87.50%	7
strongly agree	12.50%	1
Total Respondents: 8		

Q4 Should Radiology Diagnostic Accuracy Trials explicitly state that the aim is to compare index test with reference standard for diagnosis of a specific condition?

Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	62.50%	5
strongly agree	37.50%	3
Total Respondents: 8		

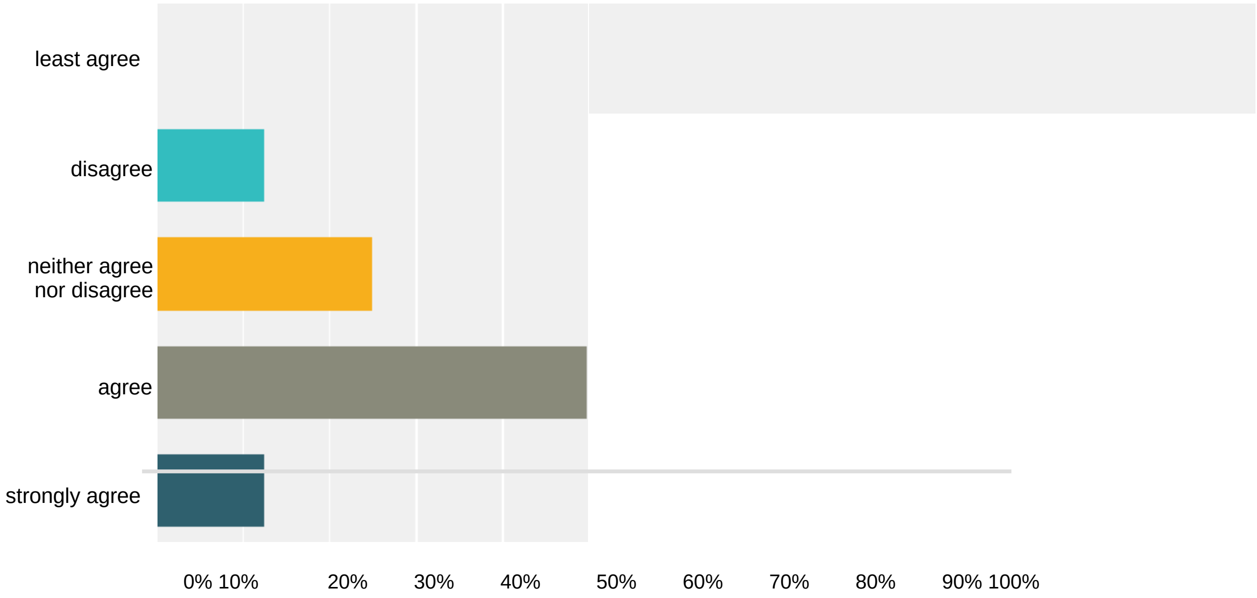
Q5 What should be included in the Title and Abstract? Please provide comment.

Answered: 4 Skipped: 4

#	Responses	Date
1	Prevalence of the disease Impact in management	11/18/2014 4:10 PM
2	At least : accuracy (or equivalent term)and both tested method(s) and used gold standard	11/8/2014 5:22 PM
3	Depends on the study, but "Diagnostic Accuracy" is usually appropriate.	10/29/2014 12:28 AM
4	The purpose of the study	10/25/2014 12:03 PM

Q6 Does this need to be included in a checklist for Radiology Diagnostic Accuracy?

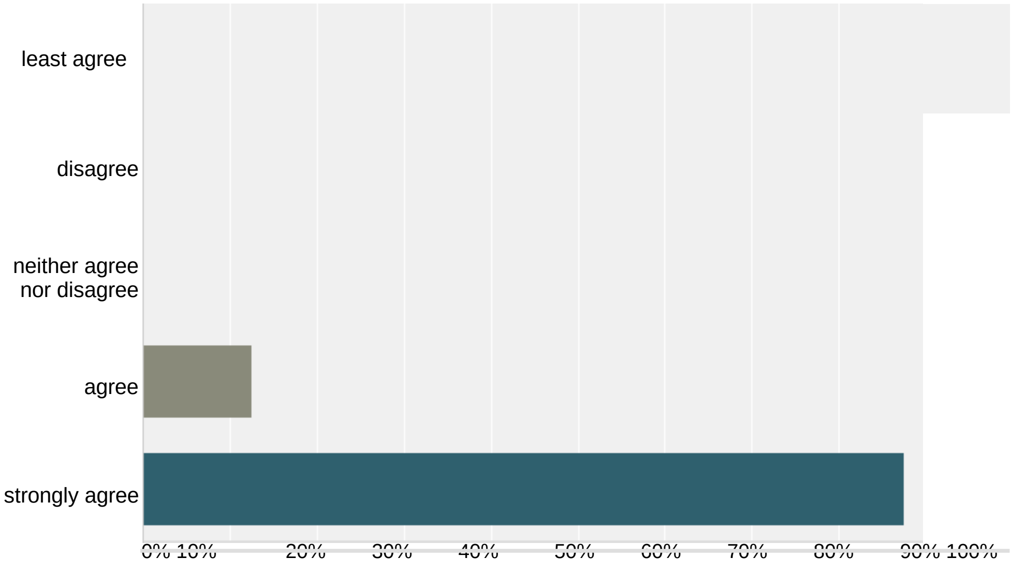
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	25.00%	2
agree	50.00%	4
strongly agree	12.50%	1
Total Respondents: 8		

Q7 Should Diagnostic Accuracy Trials provide Inclusion and Exclusion criteria for patients?

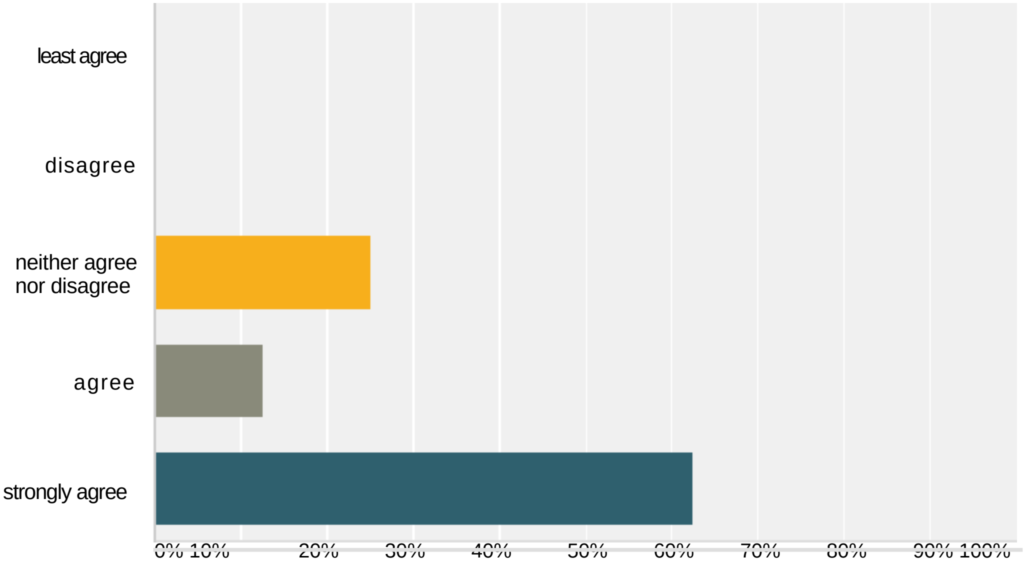
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	12.50%	1
strongly agree	87.50%	7
Total Respondents: 8		

Q8 Details of setting and location of study (e.g. whether primary or secondary care)?

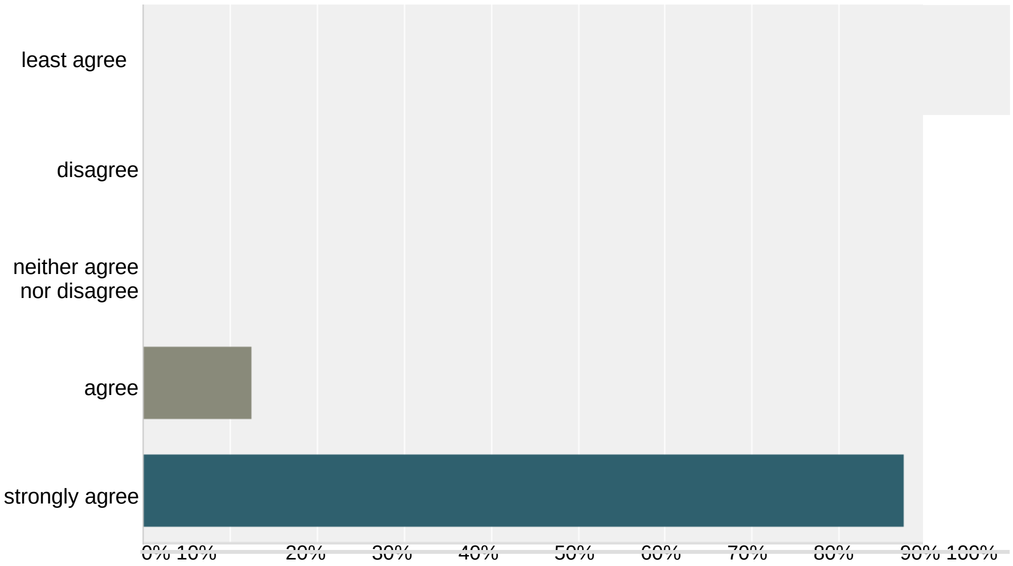
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	12.50%	1
strongly agree	62.50%	5
Total Respondents: 8		

Q9 Whether data collection is prospective or retrospective?

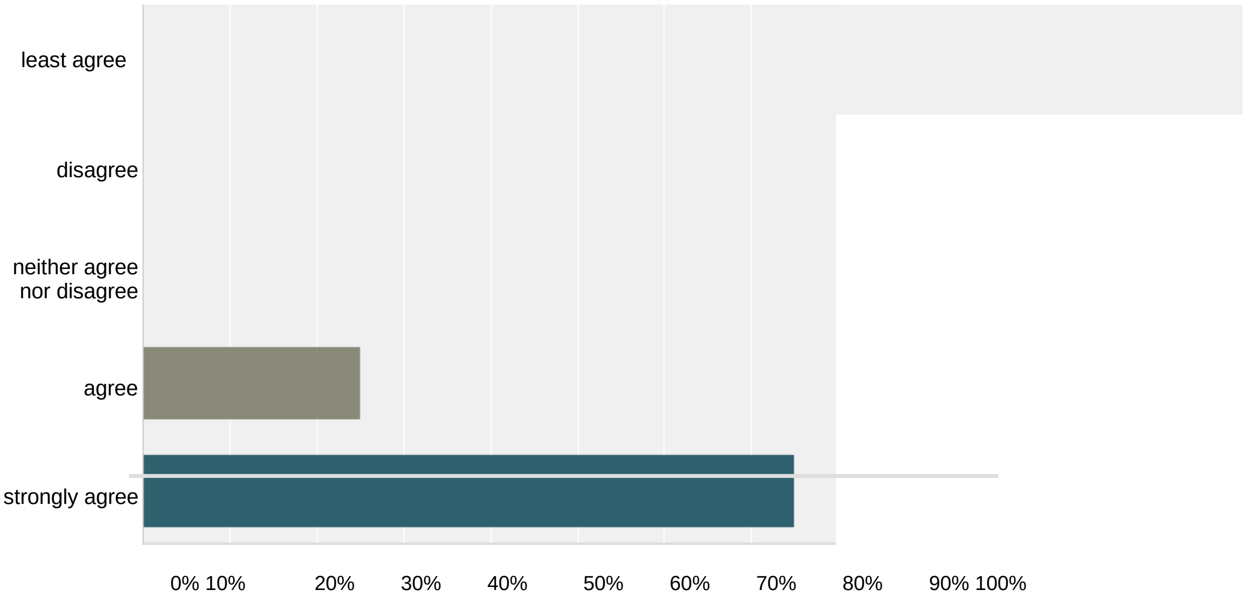
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	12.50%	1
strongly agree	87.50%	7
Total Respondents: 8		

Q10 Whether patient selection was consecutive or non-consecutive?

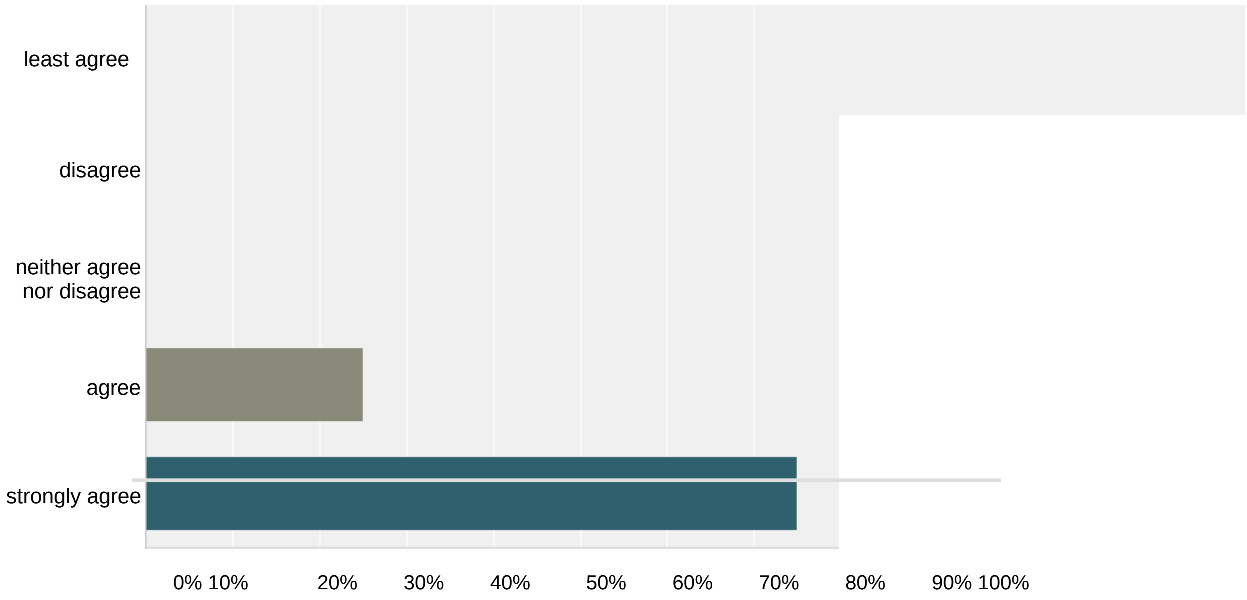
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	25.00%	2
strongly agree	75.00%	6
Total Respondents: 8		

Q11 Sample size and limitations?

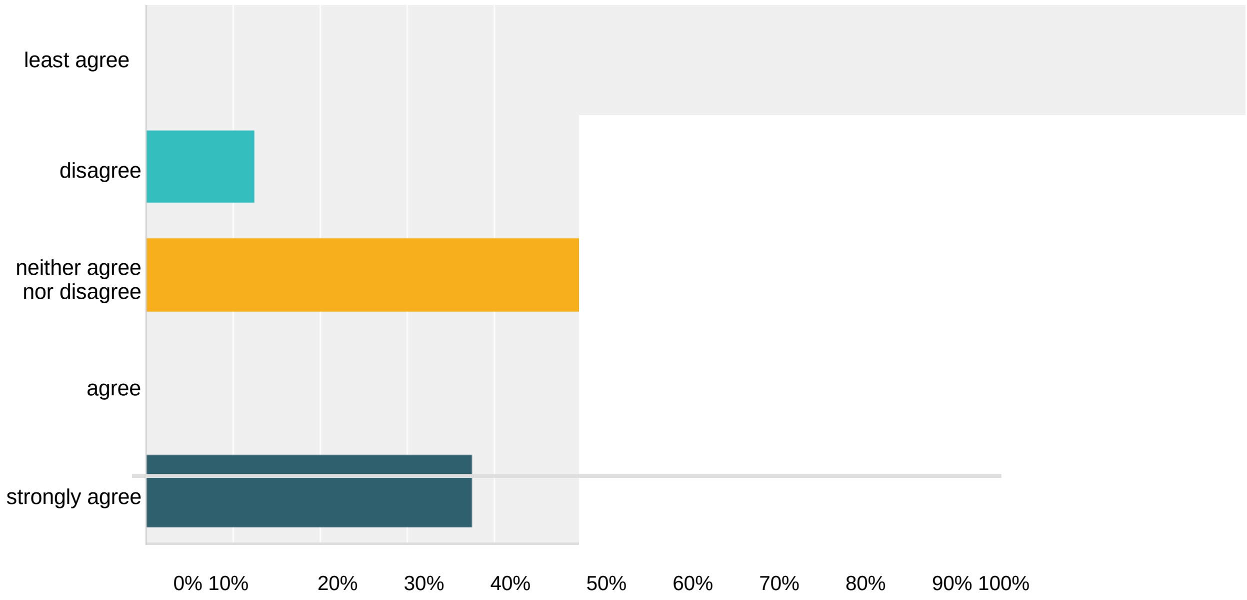
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	25.00%	2
strongly agree	75.00%	6
Total Respondents: 8		

Q12 Whether a power calculation was performed?

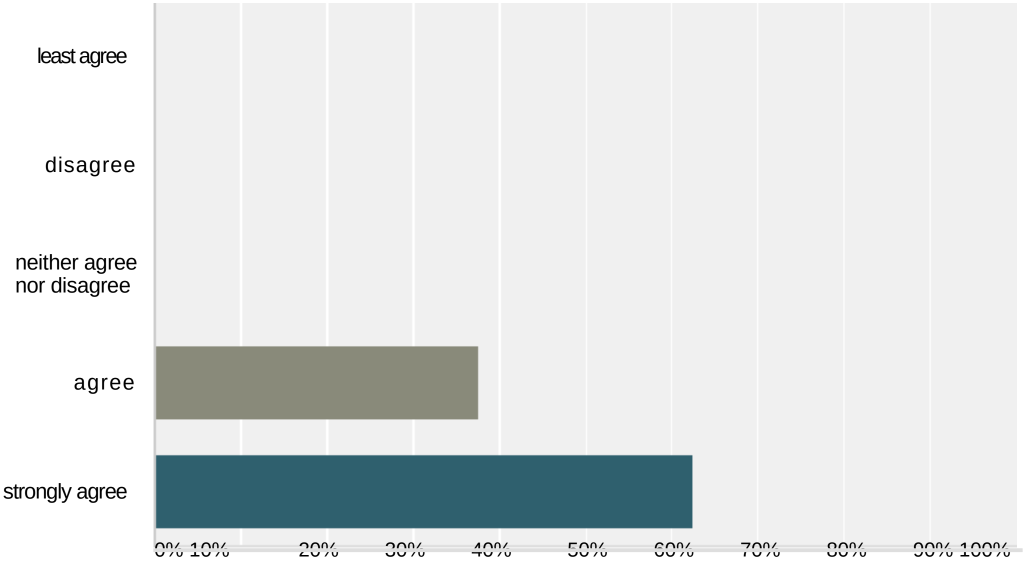
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	50.00%	4
agree	0.00%	0
strongly agree	37.50%	3
Total Respondents: 8		

Q13 Start and end dates of the study?

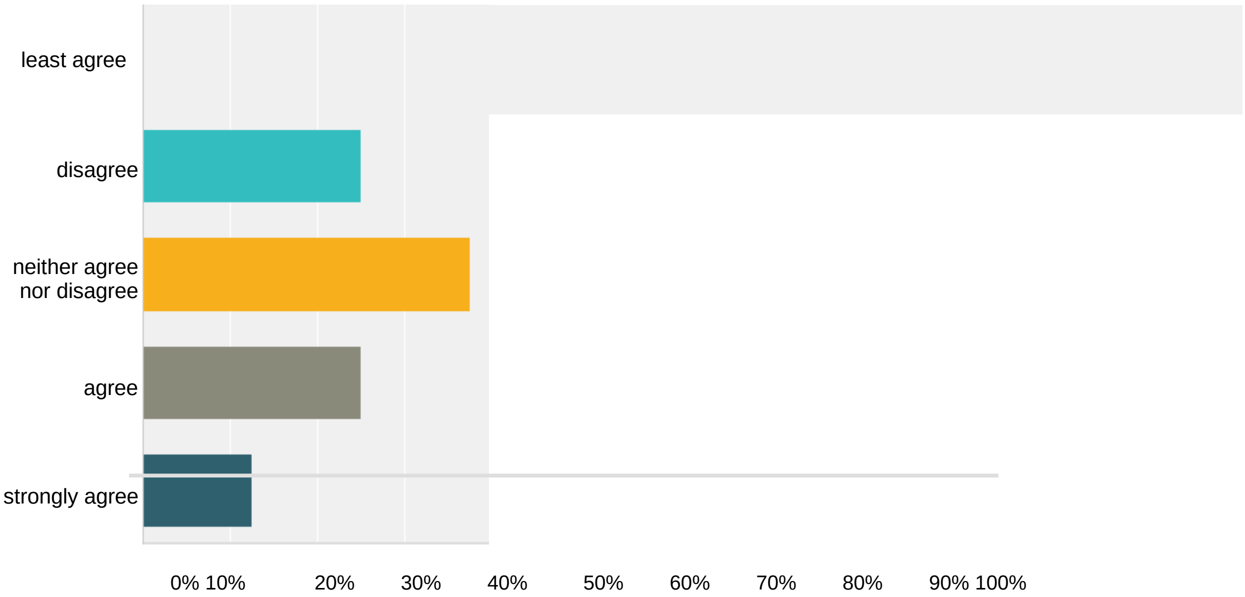
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	37.50%	3
strongly agree	62.50%	5
Total Respondents: 8		

Q14 Should a thorough description of the nature of the disease be presented?

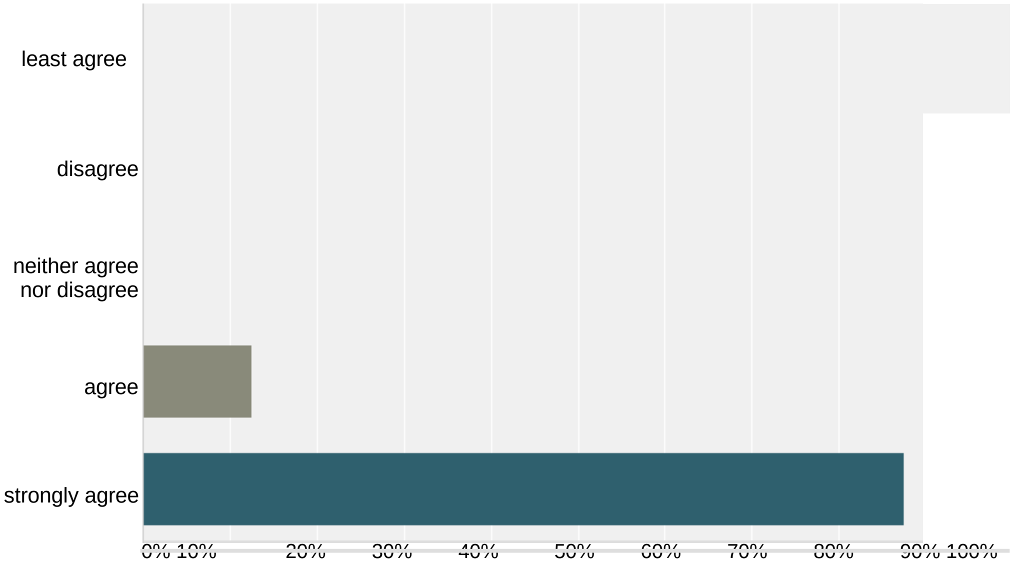
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	25.00%	2
neither agree nor disagree	37.50%	3
agree	25.00%	2
strongly agree	12.50%	1
Total Respondents: 8		

Q15 Should Radiology Diagnostic Accuracy Trials explicitly state the Reference Standard and its rationale?

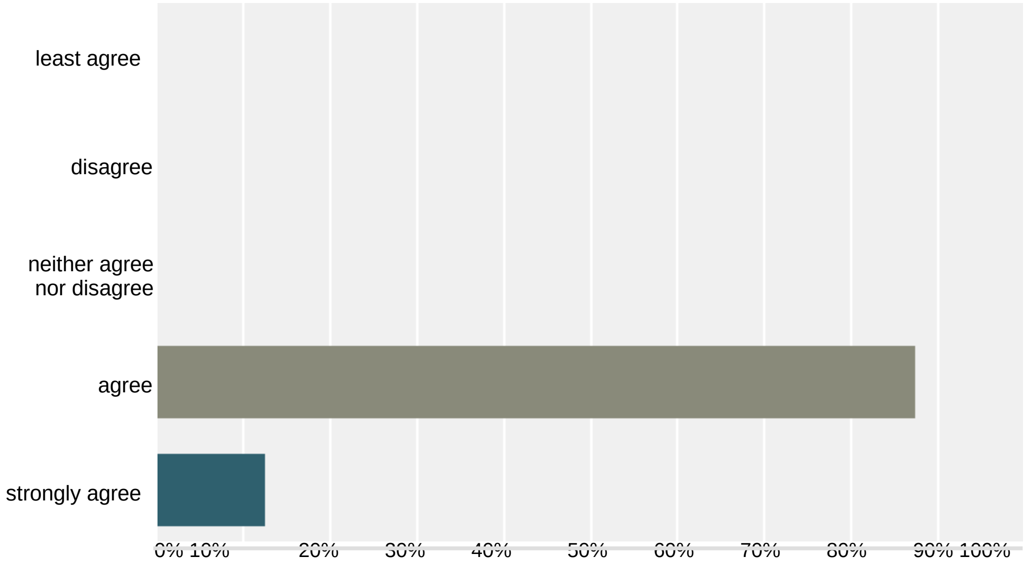
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	12.50%	1
strongly agree	87.50%	7
Total Respondents: 8		

Q16 If the Reference Standard is unavailable or imperfect, should trials state what alternative was used and justify the alternative? For example: New MRI technique for rotator cuff tear (index test) compared to arthroscopy (reference standard). If only 10% went on to surgery this would create a verification bias. Test only works in patients with severe disease creating a verification bias. Therefore would a mixed standard of reference be required if several orthopedic surgeons think there is no rotator cuff injury?

Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	87.50%	7
strongly agree	12.50%	1
Total Respondents: 8		

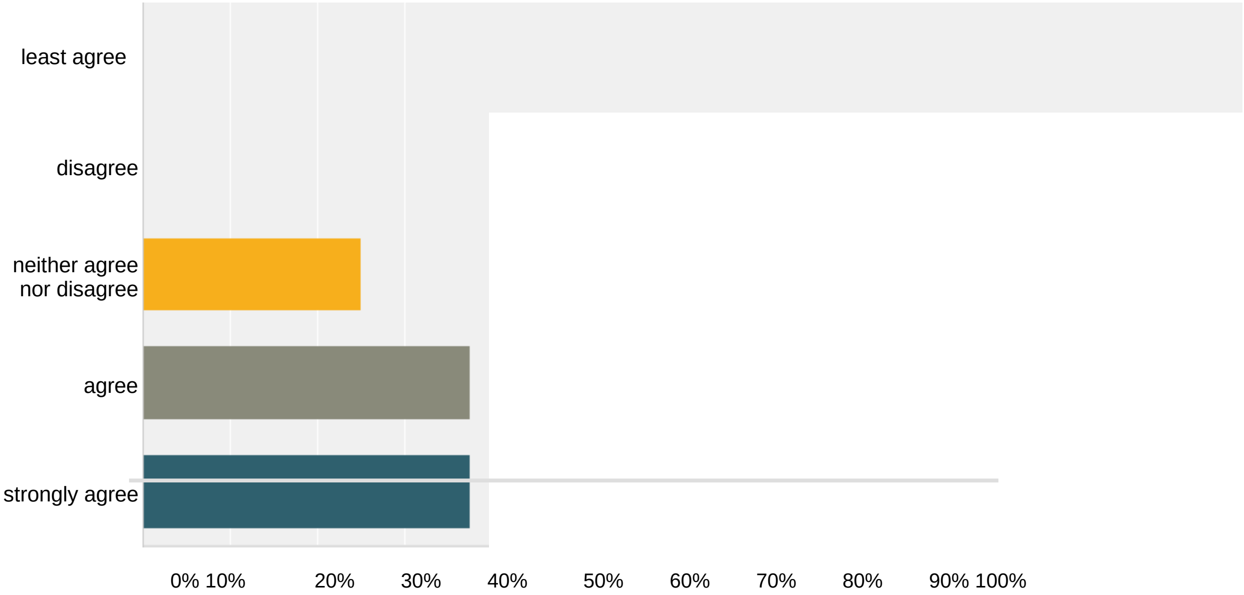
Q17 If you do not think a mixed standard or panel standard would be of benefit, please insert your comments of what would be a viable alternative for diagnostic studies with an imperfect reference standard? Please provide comment.

Answered: 1 Skipped: 7

#	Responses	Date
1	There is almost zero ways to have a perfect reference standard, and radiological studies will mostly be biased by selection . Still, the accuracy applies to the selected population under that diagnostic question.	11/8/2014 5:22 PM

Q18 In All Diagnostic Accuracy Trials?

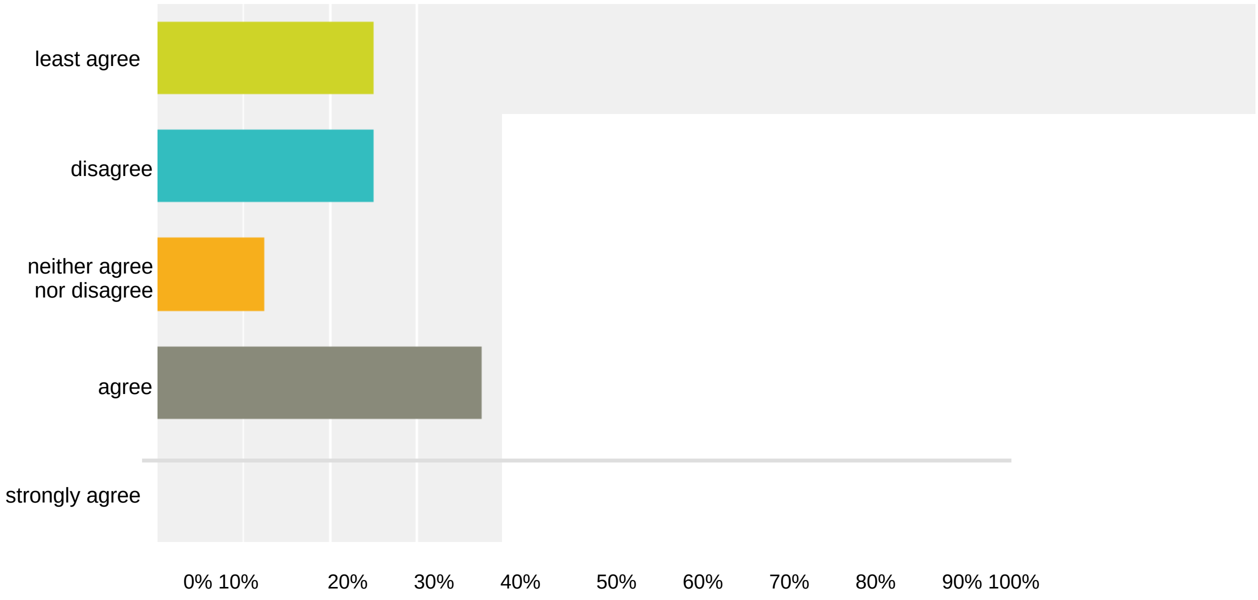
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	37.50%	3
strongly agree	37.50%	3
Total Respondents: 8		

Q19 Only where the test is not the standard of care?

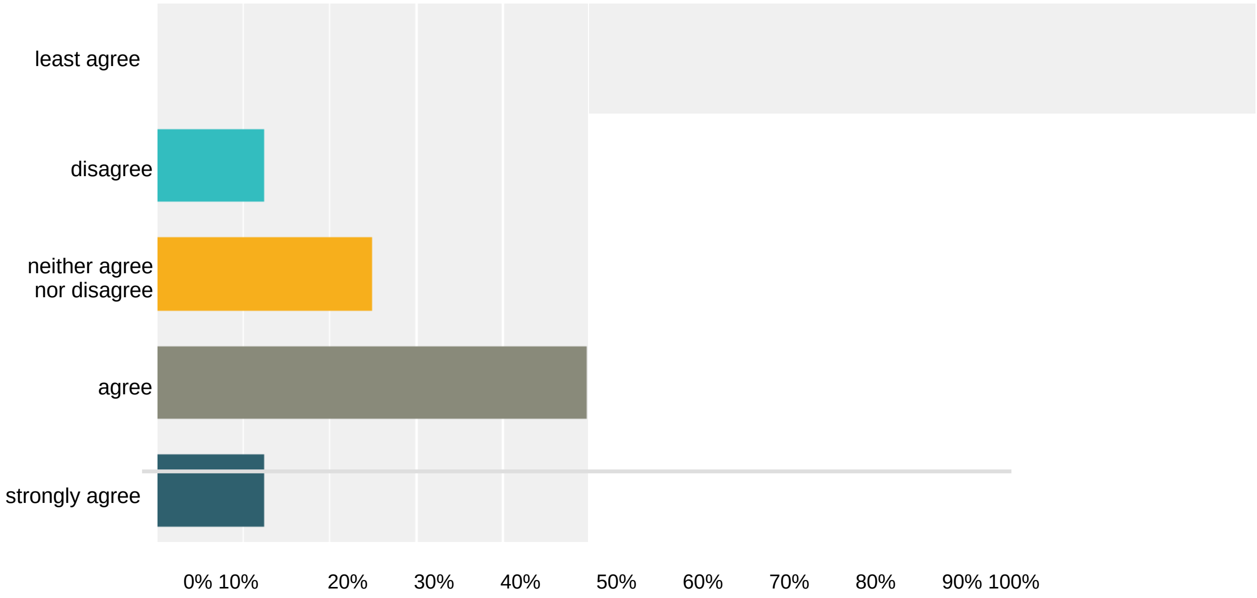
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	25.00%	2
disagree	25.00%	2
neither agree nor disagree	12.50%	1
agree	37.50%	3
strongly agree	0.00%	0
Total Respondents: 8		

Q20 Should Radiology Diagnostic Accuracy Trials explicitly state generalisability of the technique to other vendor equipment?

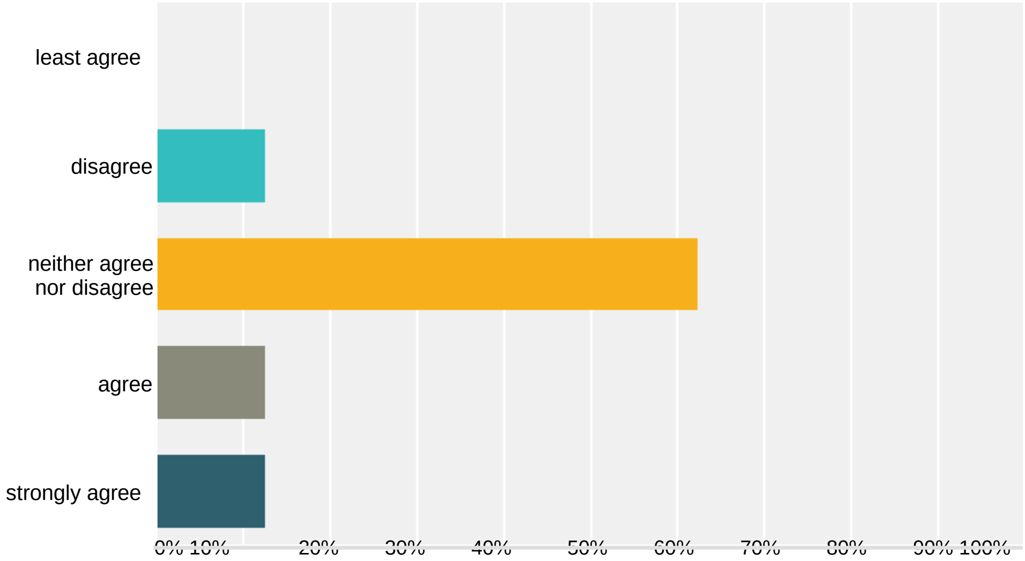
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	25.00%	2
agree	50.00%	4
strongly agree	12.50%	1
Total Respondents: 8		

Q21 For all techniques used?

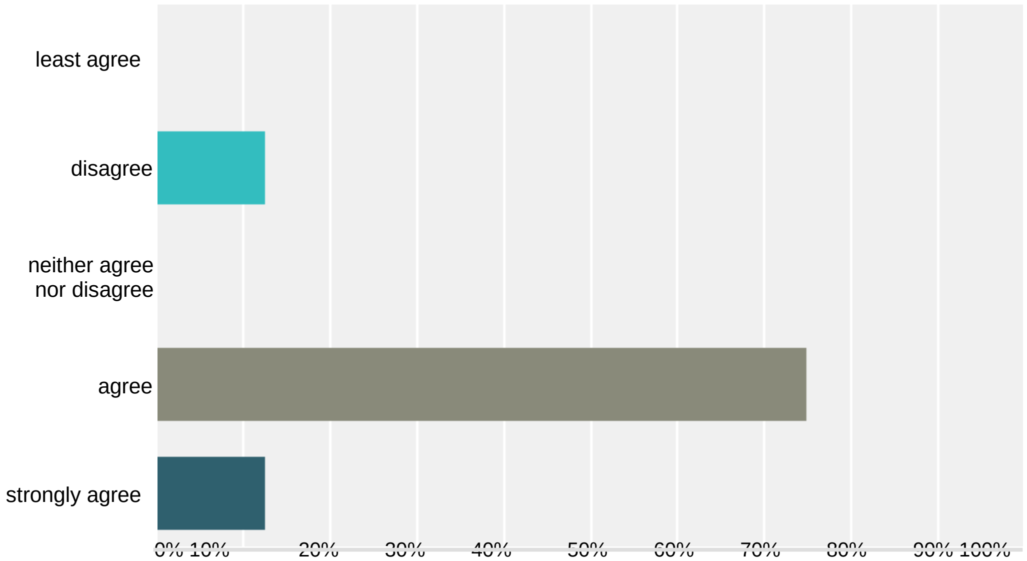
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	62.50%	5
agree	12.50%	1
strongly agree	12.50%	1
Total Respondents: 8		

Q22 Only for new techniques?

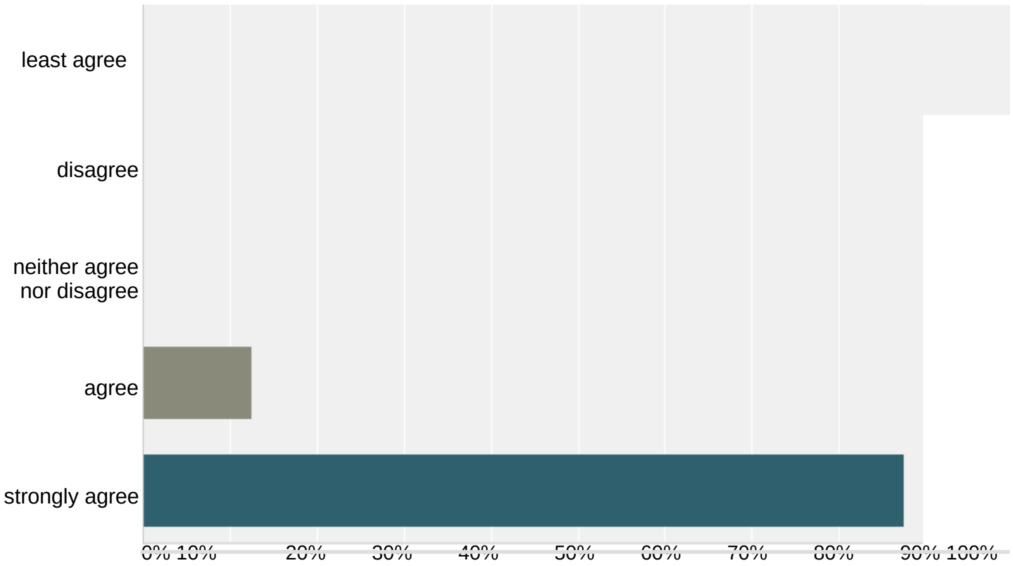
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	0.00%	0
agree	75.00%	6
strongly agree	12.50%	1
Total Respondents: 8		

Q23 If the test was modified during the study, do you need to know?

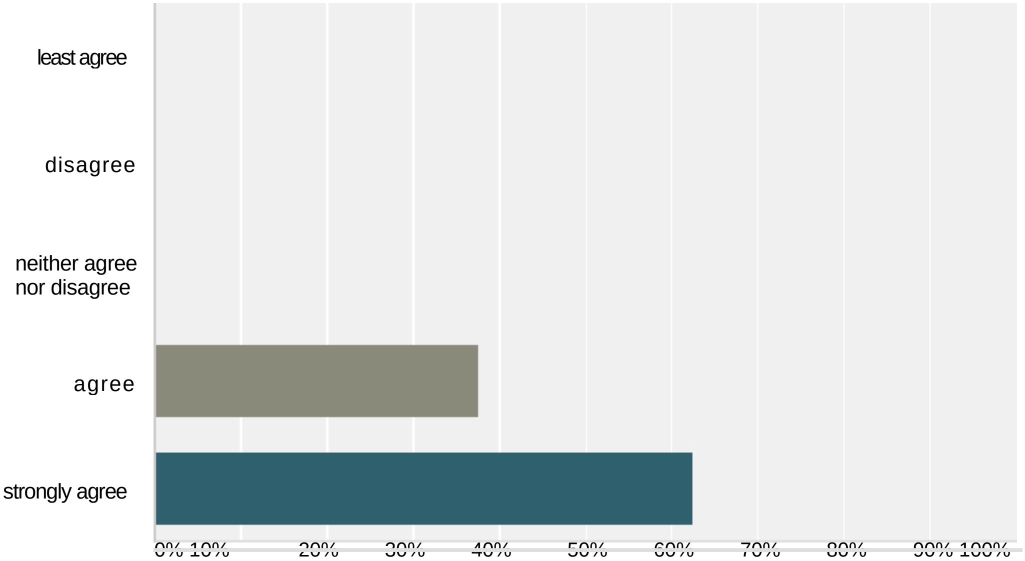
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	12.50%	1
strongly agree	87.50%	7
Total Respondents: 8		

Q24 Should Radiology Diagnostic Accuracy Trials report cut-off values or specific diagnostic criteria for index and reference tests?

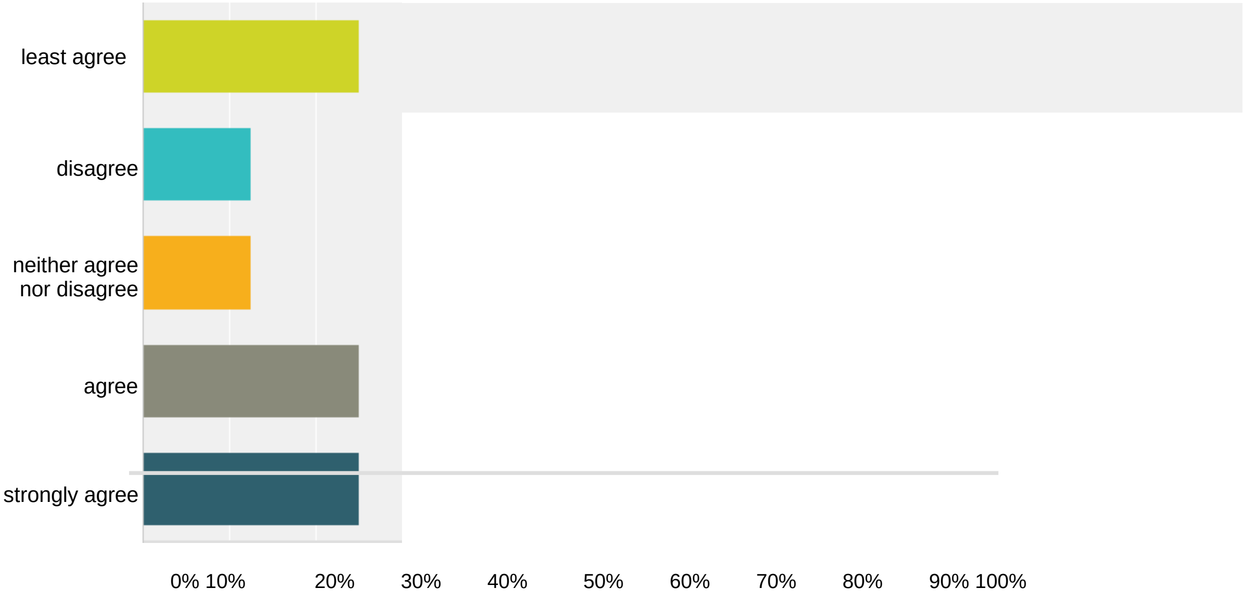
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	37.50%	3
strongly agree	62.50%	5
Total Respondents: 8		

Q25 Should Radiology Diagnostic Accuracy Trials provide robust assessment of inter-observer agreement including >3 observers with variable expertise?

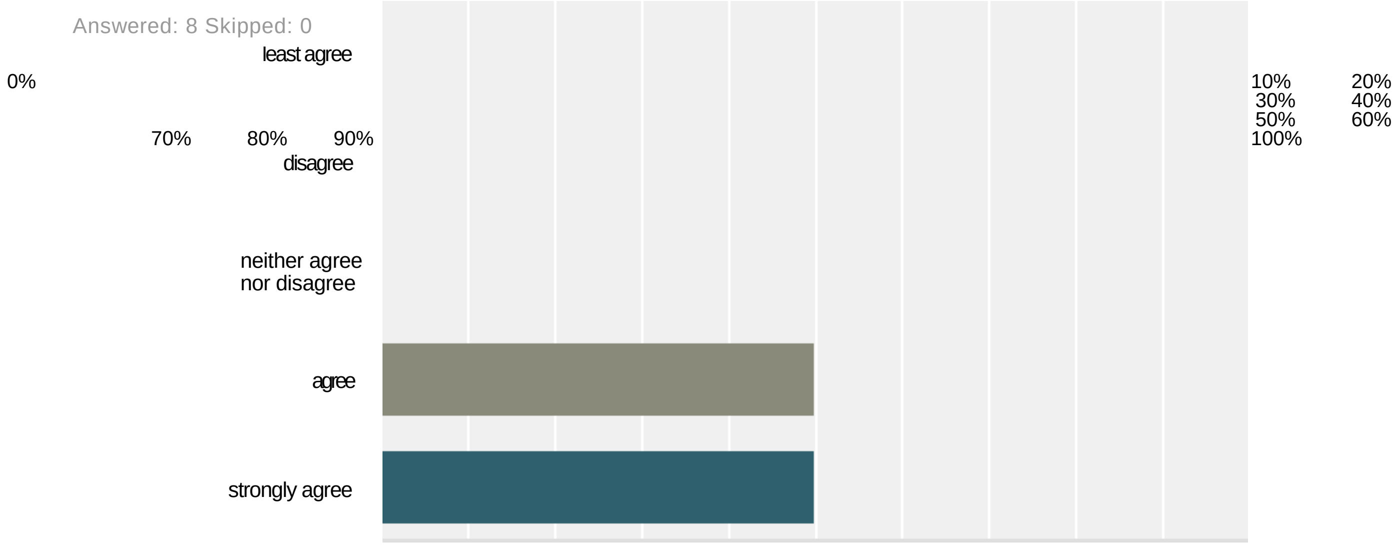
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	25.00%	2
disagree	12.50%	1
neither agree nor disagree	12.50%	1
agree	25.00%	2
strongly agree	25.00%	2
Total Respondents: 8		

Q26 Training and number of investigators?

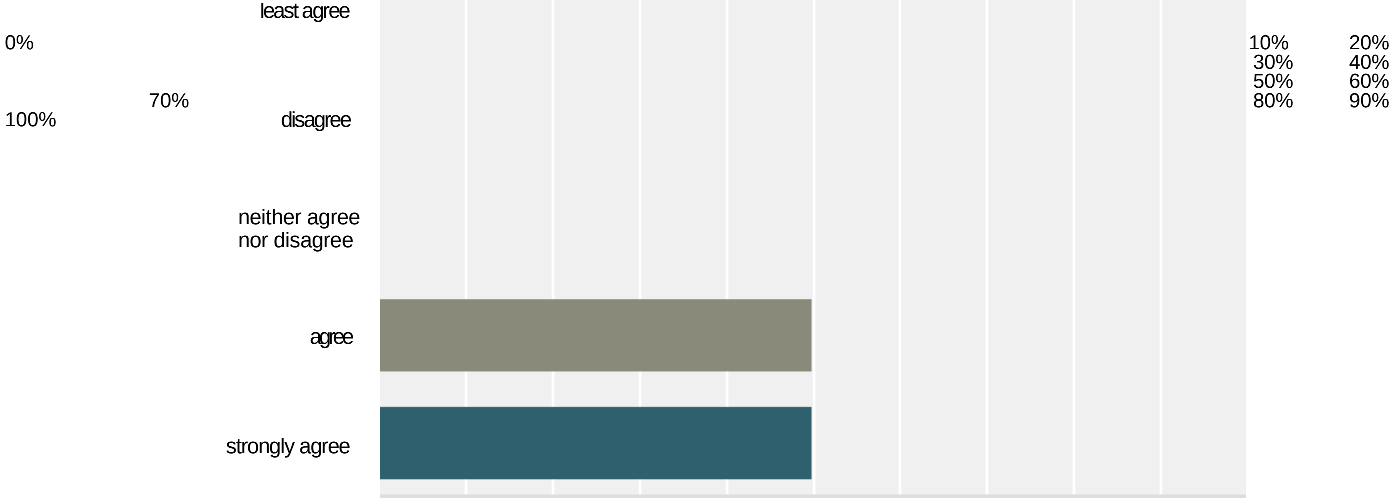
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	50.00%	4
strongly agree	50.00%	4
Total Respondents: 8		

Q27 Whether extra training was provided for a new technique?

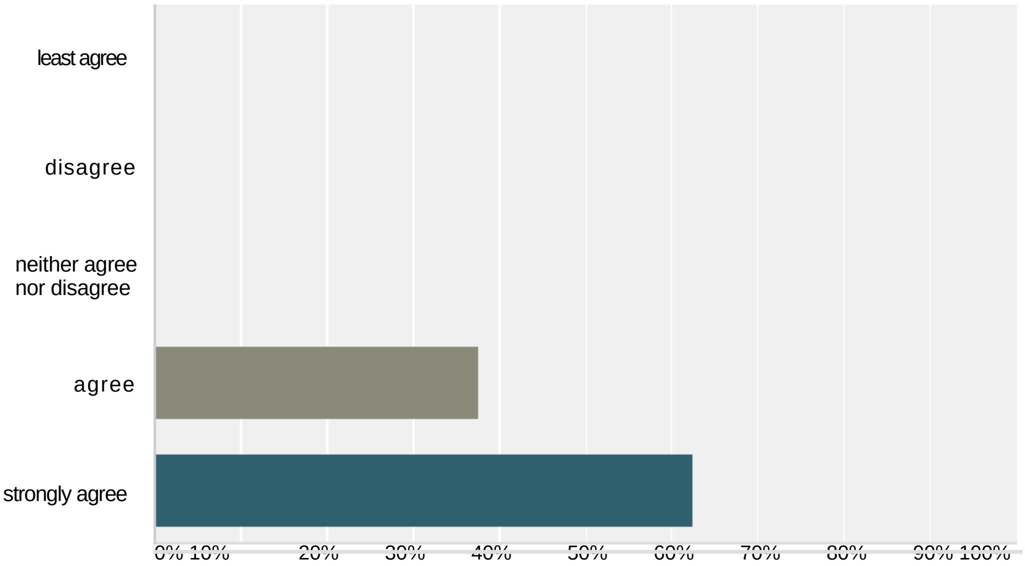
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	50.00%	4
strongly agree	50.00%	4
Total Respondents: 8		

Q28 Whether readers were blinded to prior test results?

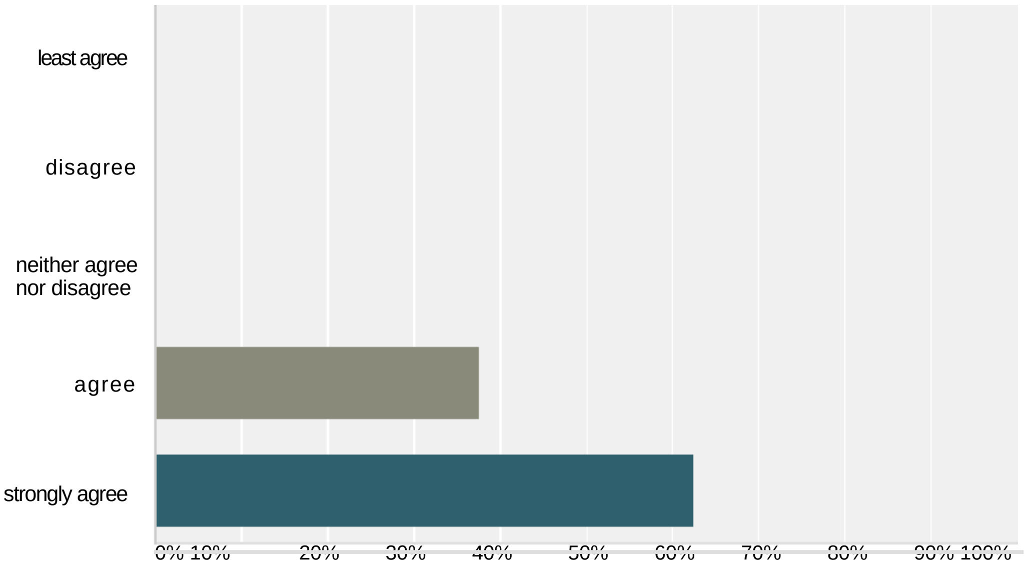
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	37.50%	3
strongly agree	62.50%	5
Total Respondents: 8		

Q29 Whether readers were blinded to clinical information?

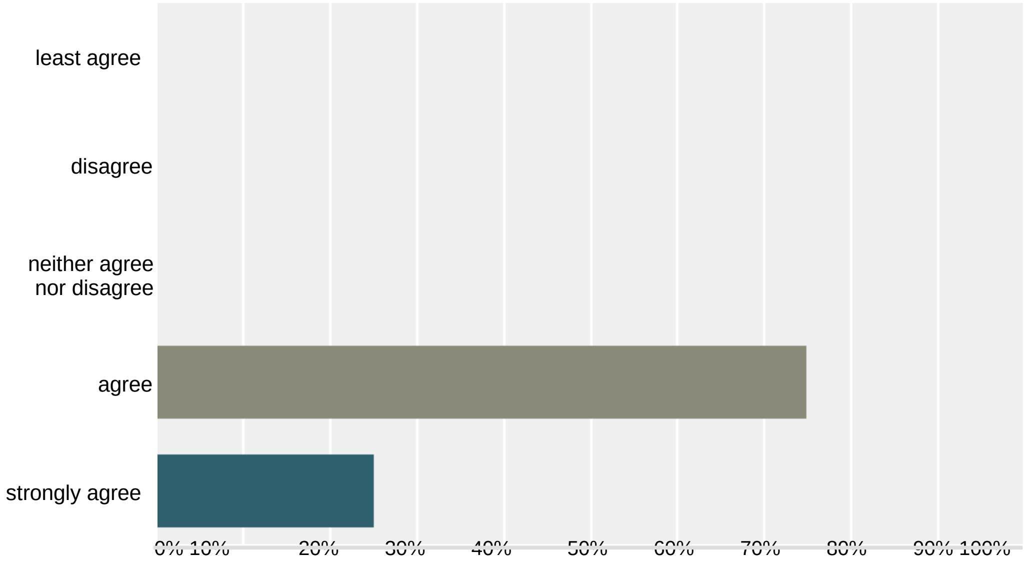
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	37.50%	3
strongly agree	62.50%	5
Total Respondents: 8		

Q30 Should Radiology Diagnostic Accuracy Trials report study flow, including eligible patients who did not undergo index or reference tests, and explain why?

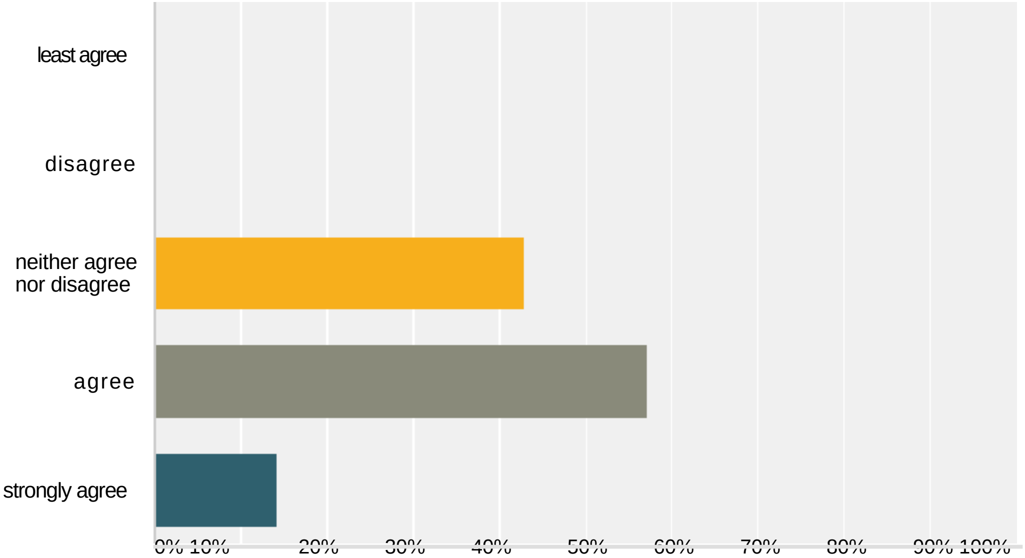
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	75.00%	6
strongly agree	25.00%	2
Total Respondents: 8		

Q31 Should a flow diagram be provided?

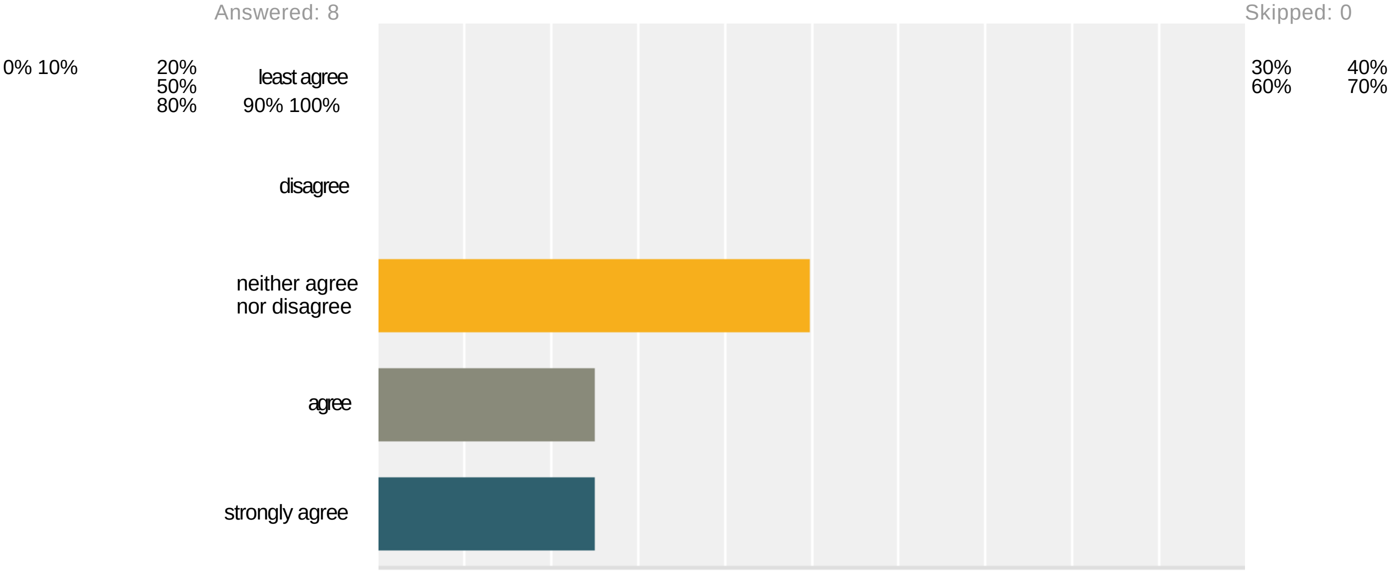
Answered: 7 Skipped: 1



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	42.86%	3
agree	57.14%	4
strongly agree	14.29%	1
Total Respondents: 7		

Q32 Should changes in diagnosis after each test be reported?

Answered: 8 Skipped: 0



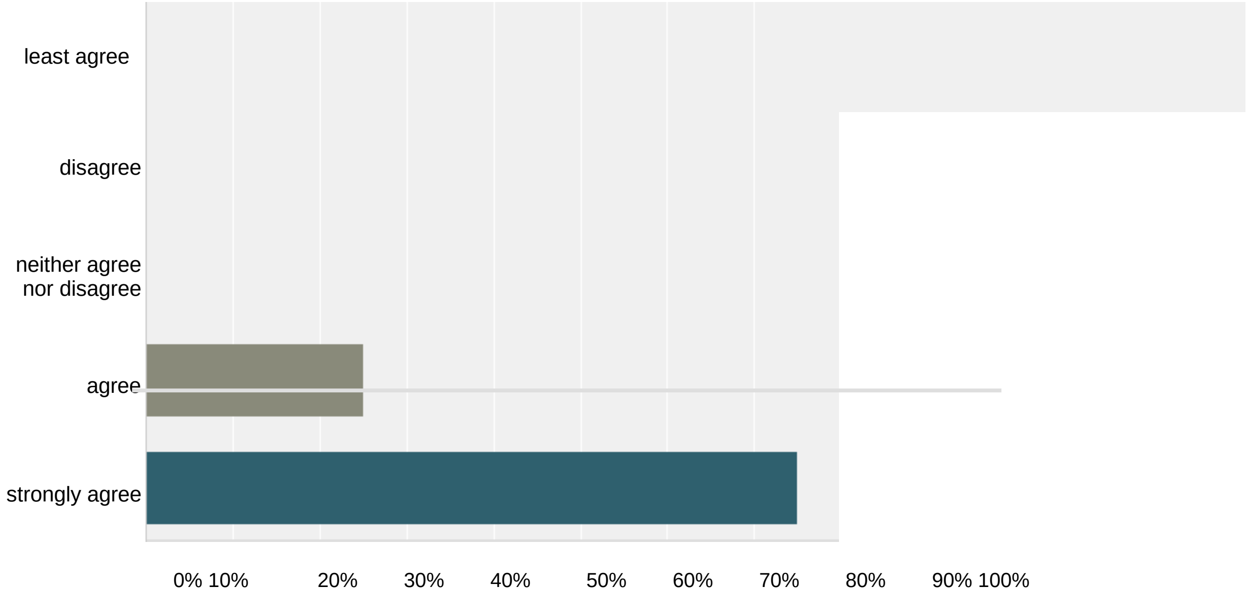
Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	50.00%	4
agree	25.00%	2
strongly agree	25.00%	2
Total Respondents: 8		

Q33 Should patient demographics provided including age, sex, presenting diagnosis or

symptoms, any co-morbidities, and

concurrent therapies be provided?

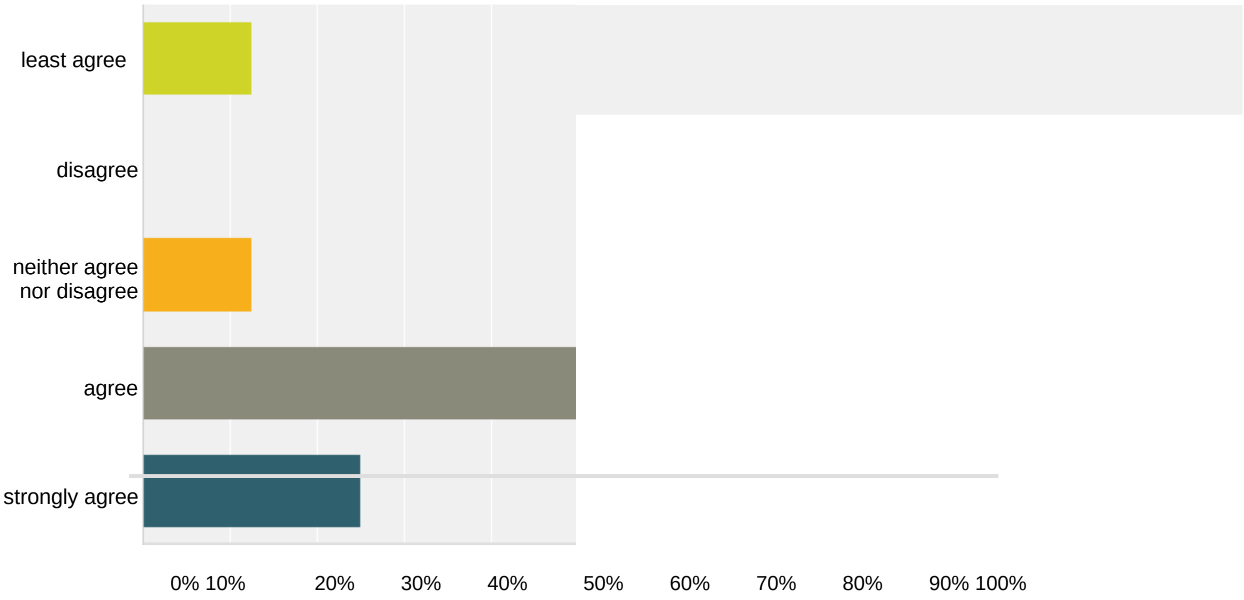
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	25.00%	2
strongly agree	75.00%	6
Total Respondents: 8		

Q34 Should the severity (spectrum) of the disease entity be explicitly reported?

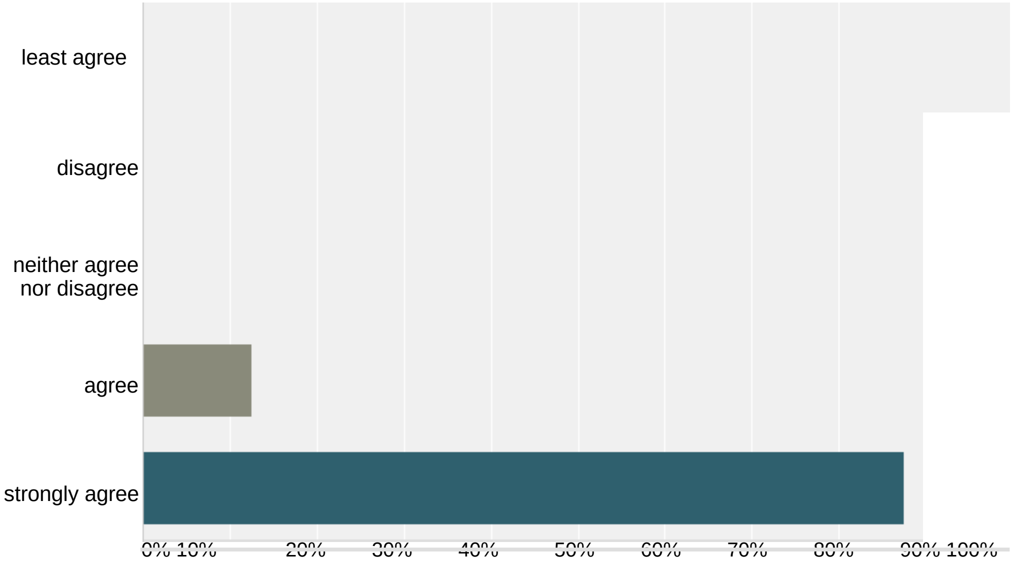
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	12.50%	1
agree	50.00%	4
strongly agree	25.00%	2
Total Respondents: 8		

Q35 Should the time difference between the index test and reference standard be provided?

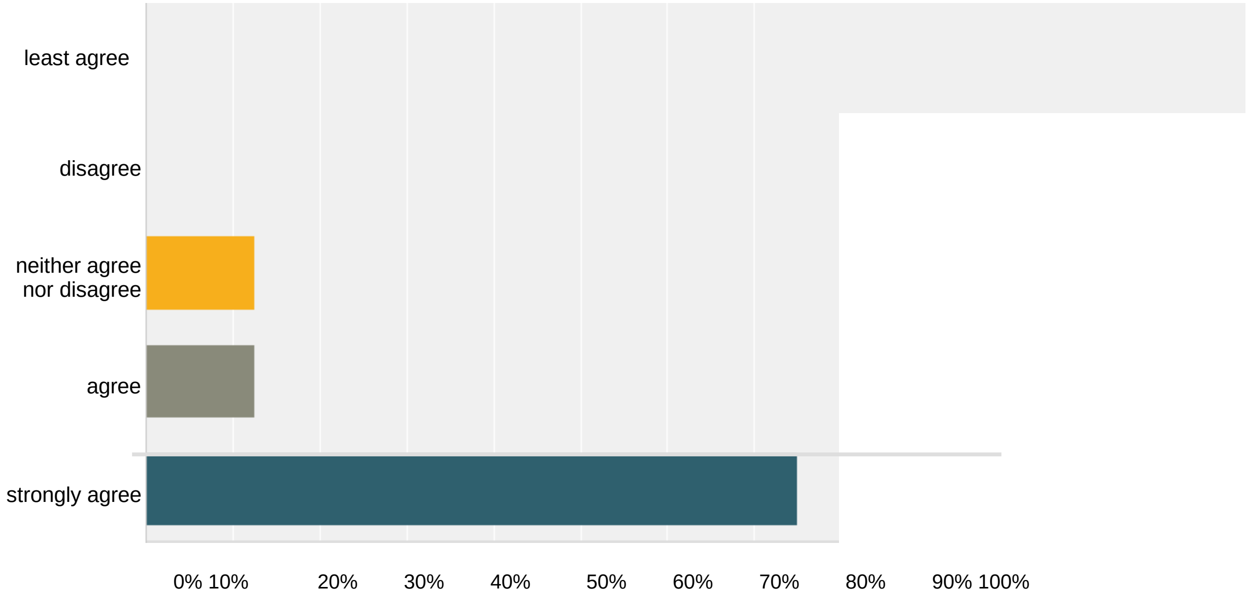
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	12.50%	1
strongly agree	87.50%	7
Total Respondents: 8		

Q36 Do you need to know if any other treatments were provided between the two tests?

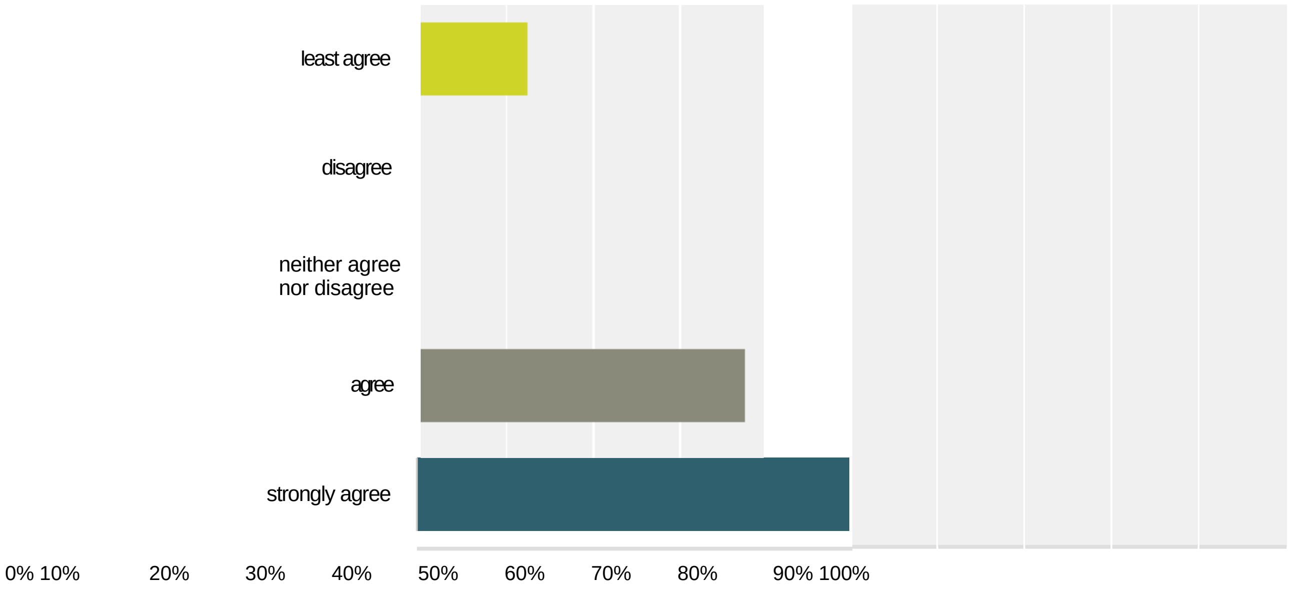
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	12.50%	1
agree	12.50%	1
strongly agree	75.00%	6
Total Respondents: 8		

Q37 Should adverse events be reported for either the index test or reference standard?

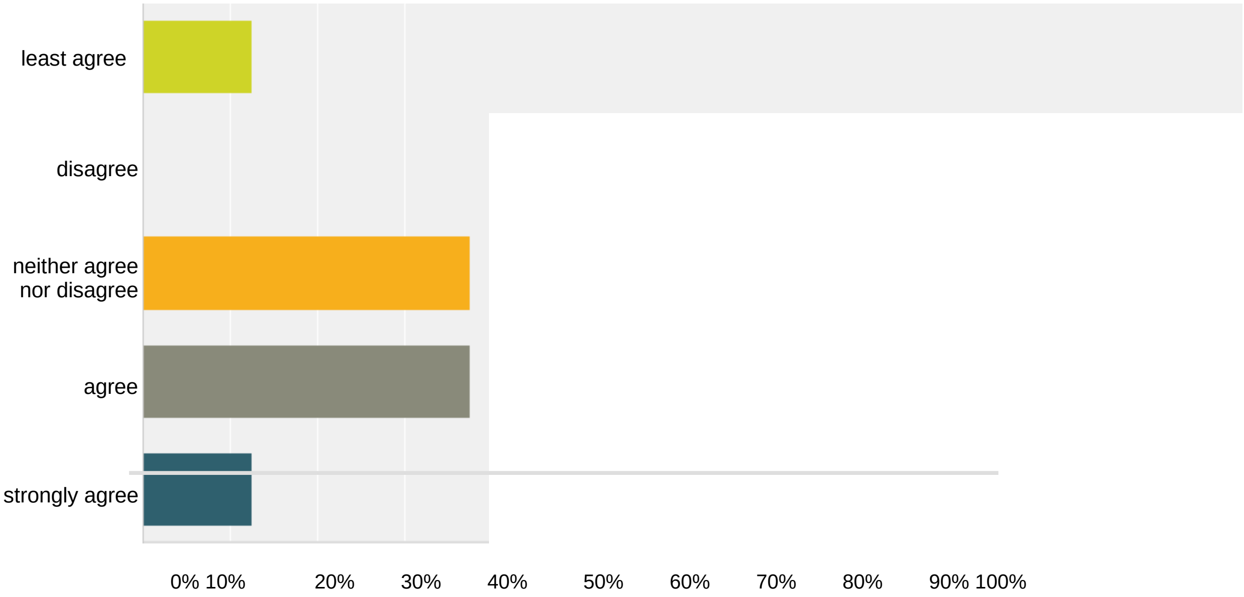
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	37.50%	3
strongly agree	50.00%	4
Total Respondents: 8		

Q38 Should it be stated if contrast could have been avoided?

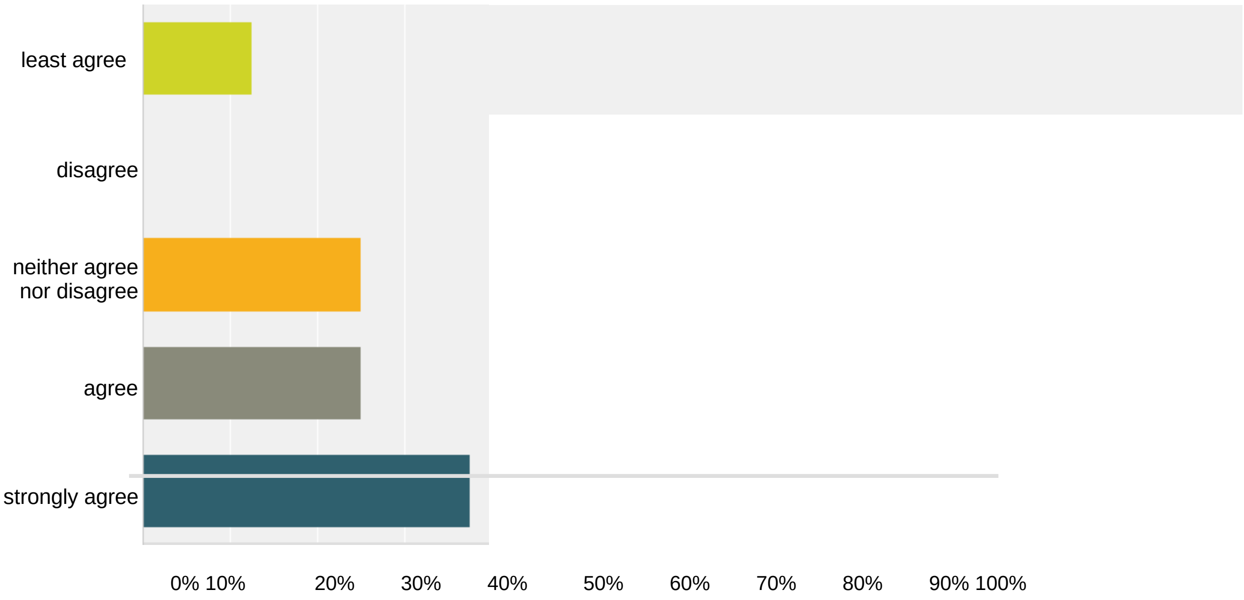
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	37.50%	3
agree	37.50%	3
strongly agree	12.50%	1
Total Respondents: 8		

Q39 Would you need to know if the level of radiation was less for the index test?

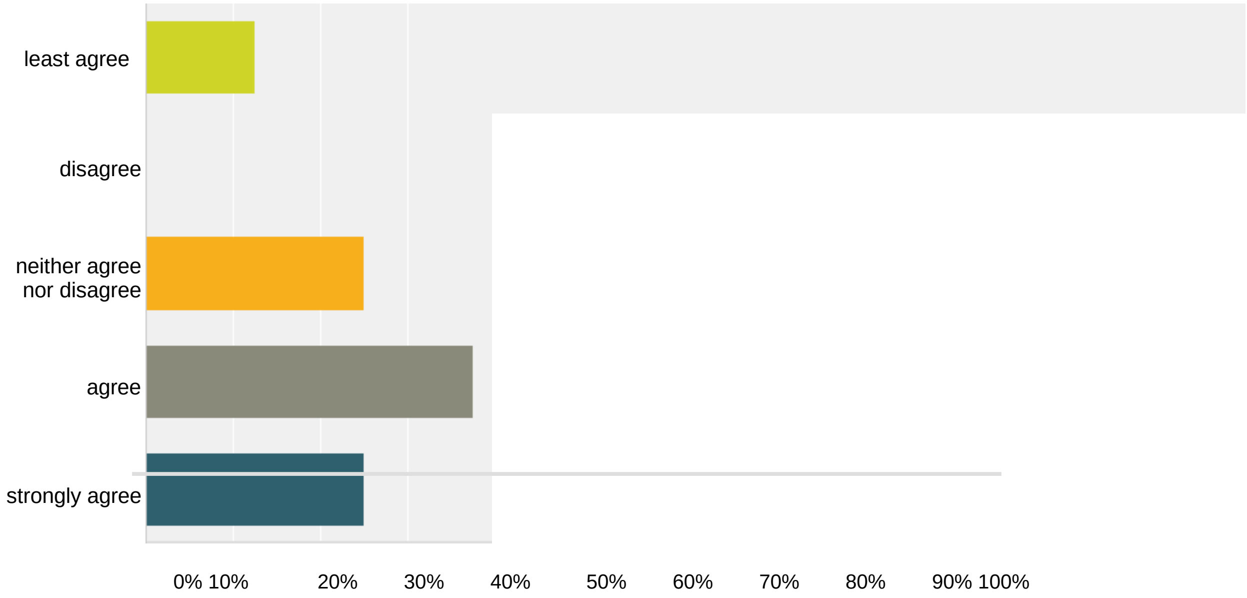
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	25.00%	2
strongly agree	37.50%	3
Total Respondents: 8		

Q40 Should Radiology Diagnostic Accuracy Trials include a cross tabulation of results of index and reference tests?

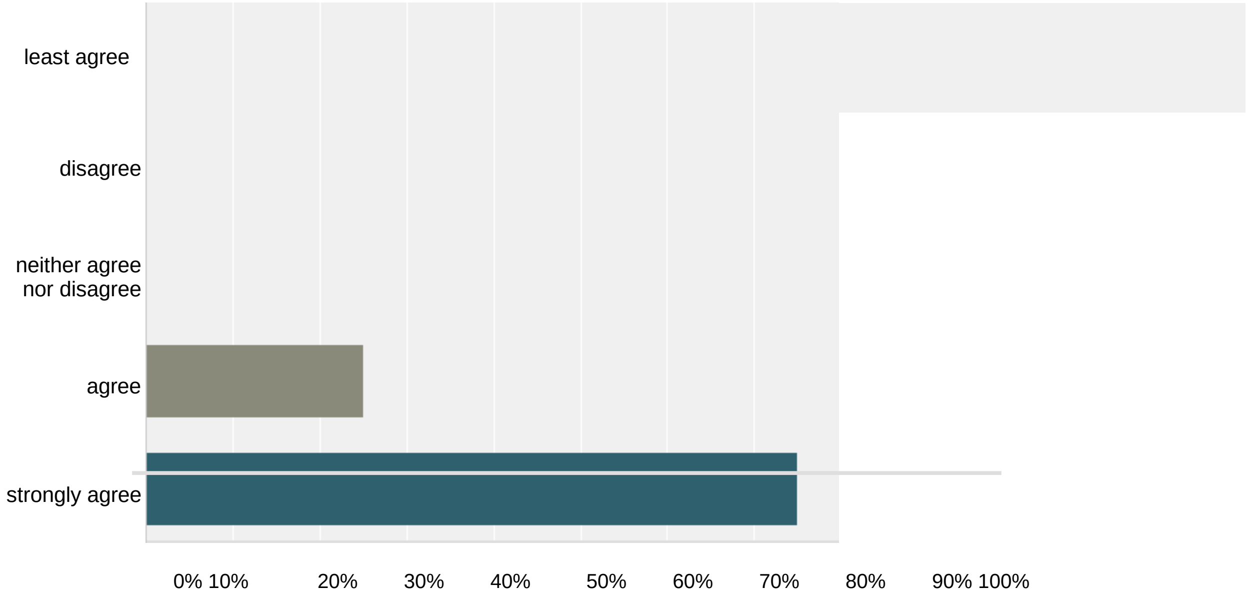
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	37.50%	3
strongly agree	25.00%	2
Total Respondents: 8		

Q41 Radiology Diagnostic Accuracy Trials should explicitly state measures of diagnostic accuracy and uncertainty (preferably p-values and 95% CI).

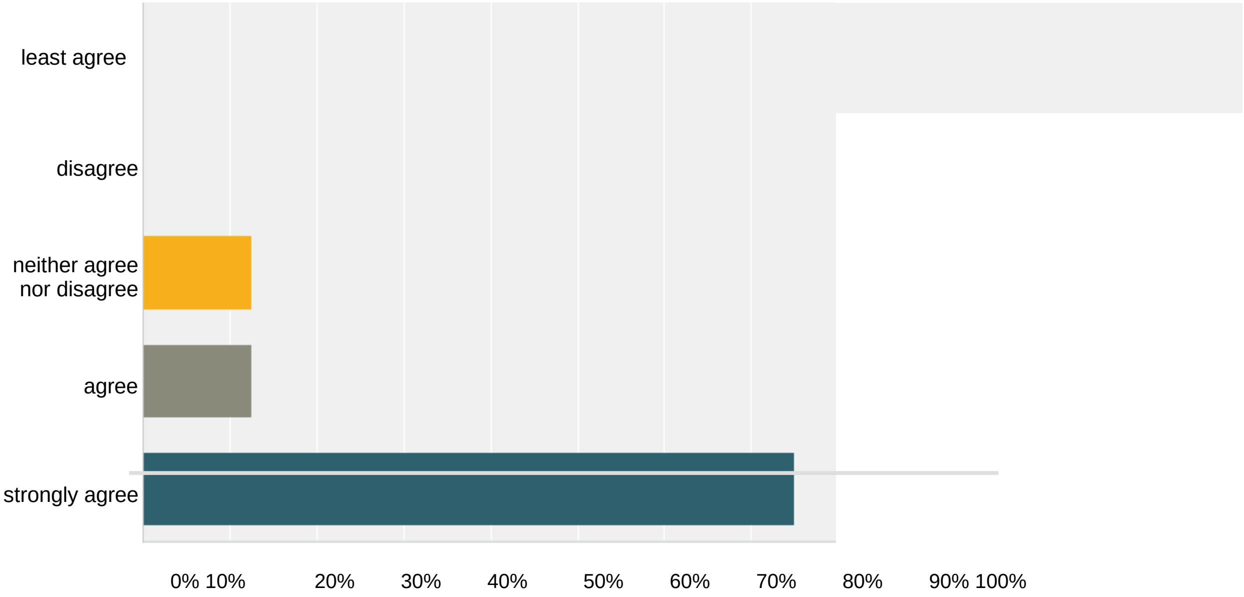
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	25.00%	2
strongly agree	75.00%	6
Total Respondents: 8		

Q42 Radiology Diagnostic Accuracy Trials should report indeterminate/missing results and outliers; and describe how this data was handled.

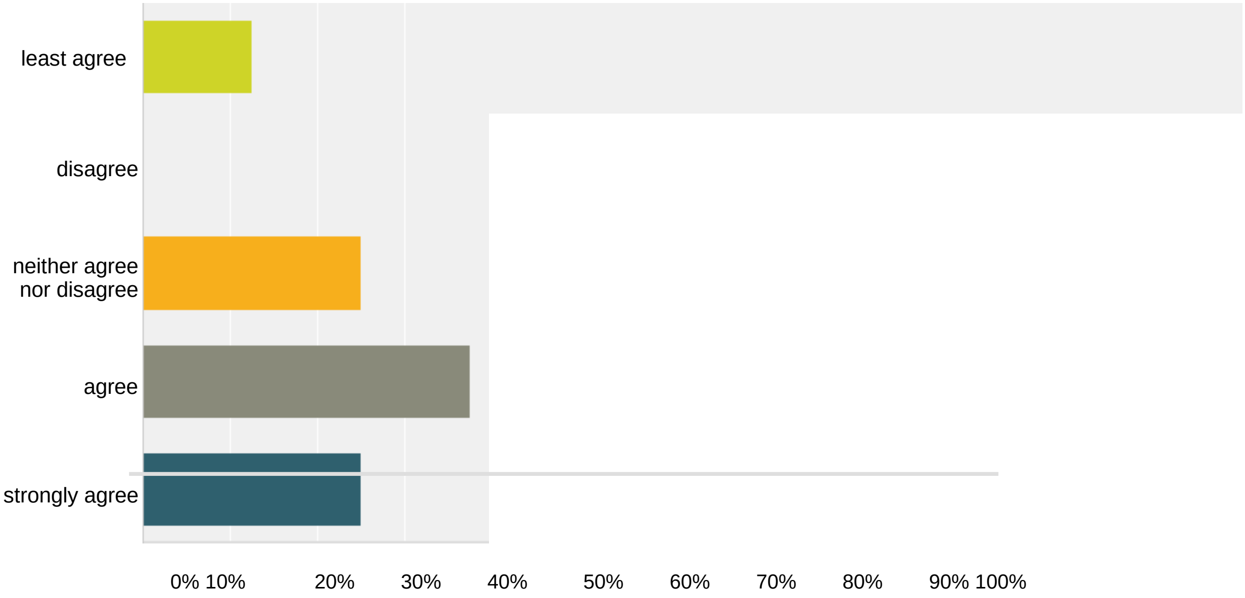
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	12.50%	1
agree	12.50%	1
strongly agree	75.00%	6
Total Respondents: 8		

Q43 Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.

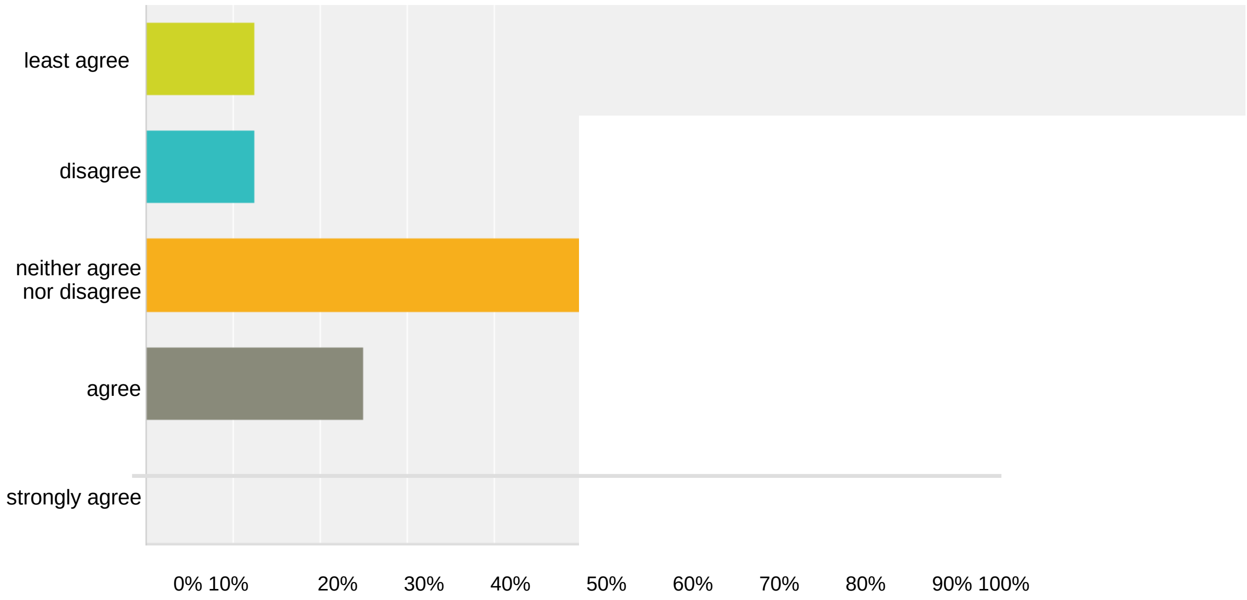
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	0.00%	0
neither agree nor disagree	25.00%	2
agree	37.50%	3
strongly agree	25.00%	2
Total Respondents: 8		

Q44 To ensure reproducibility do you think that re-analysis of the index test and reference standard should be indicated?

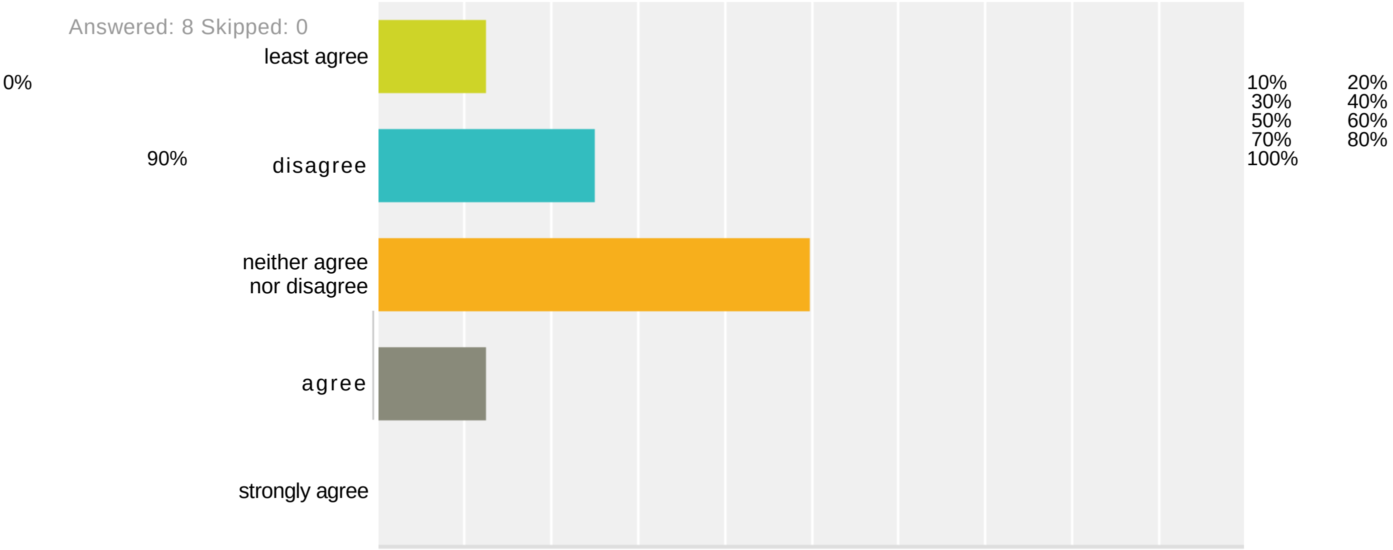
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	12.50%	1
neither agree nor disagree	50.00%	4
agree	25.00%	2
strongly agree	0.00%	0
Total Respondents: 8		

Q45 Should this be provided all the time?

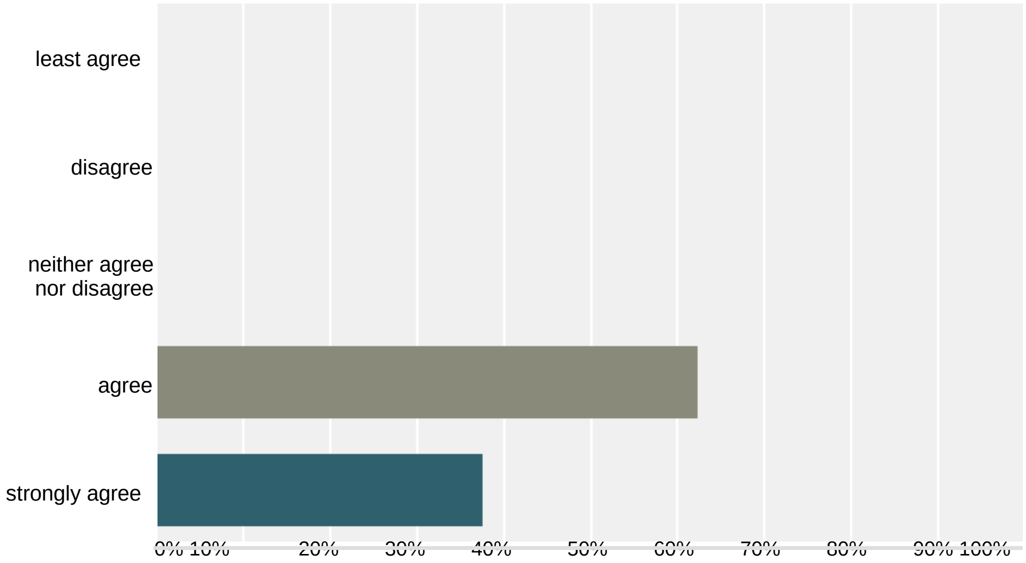
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	12.50%	1
disagree	25.00%	2
neither agree nor disagree	50.00%	4
agree	12.50%	1
strongly agree	0.00%	0
Total Respondents: 8		

Q46 Radiology Diagnostic Accuracy Trials should include estimates of inter-observer variability.

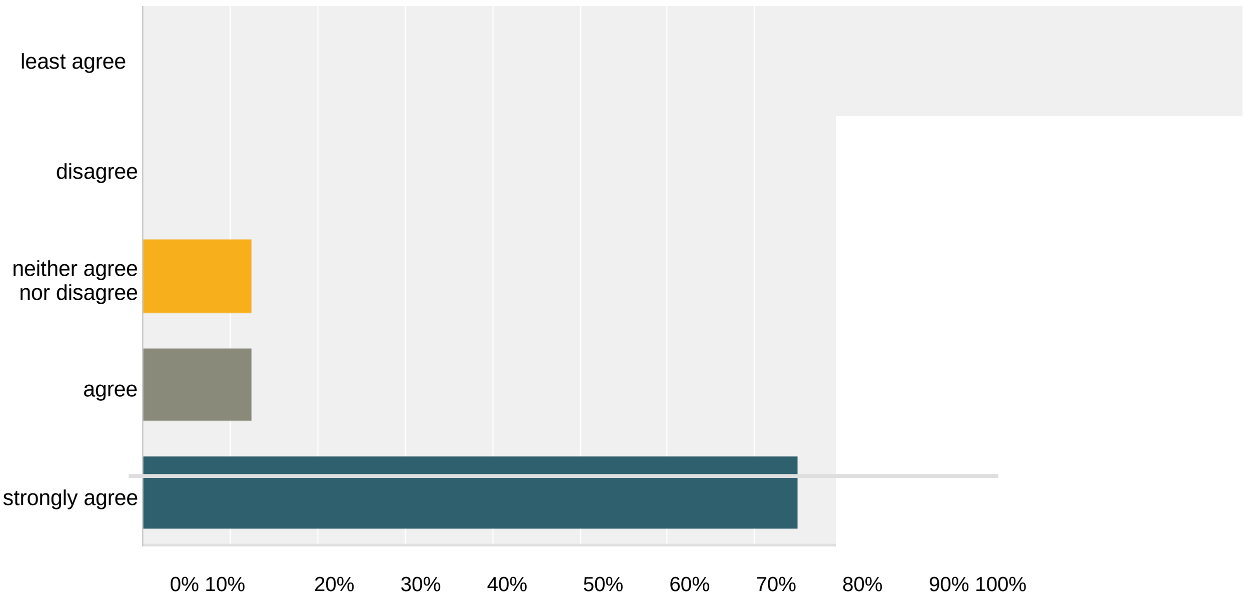
Answered: 8 Skipped: 0



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	0.00%	0
agree	62.50%	5
strongly agree	37.50%	3
Total Respondents: 8		

Q47 Should the clinical relevance of the study findings be provided?

Answered: 8 Skipped: 0



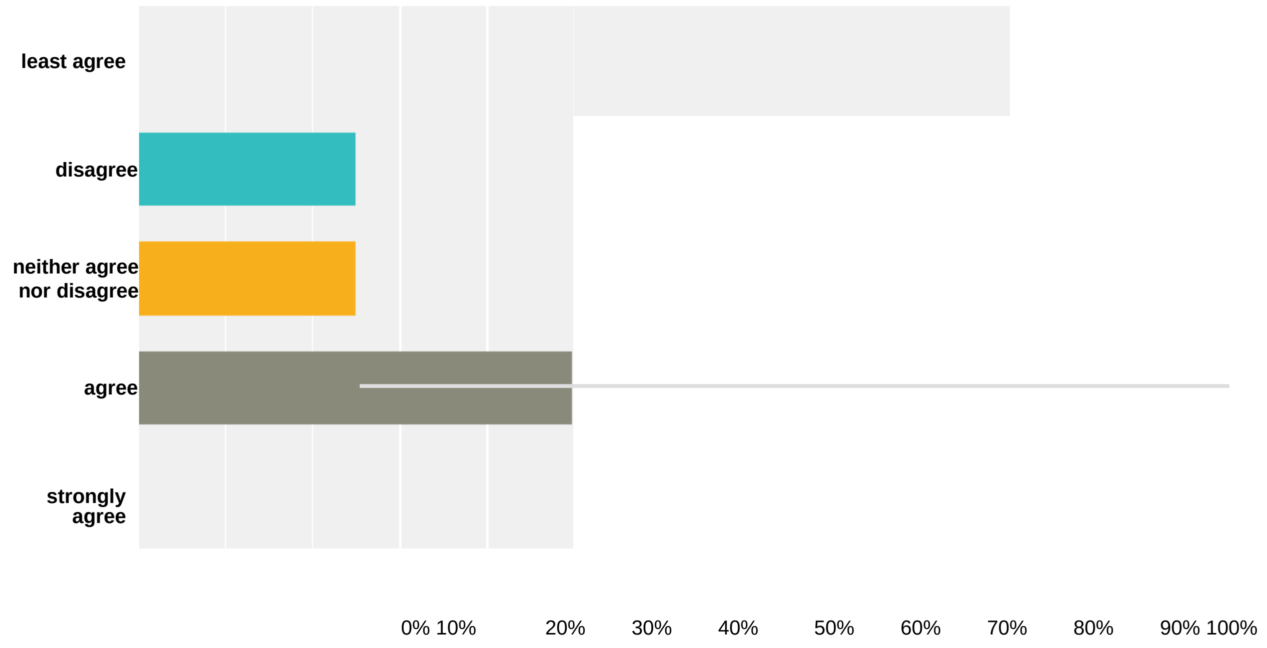
Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	12.50%	1
agree	12.50%	1
strongly agree	75.00%	6
Total Respondents: 8		

Appendix 28: Results of Delphi Round 2

STARD_DI Delphi Round 2

Q1 Whether a power calculation was performed.

Answered: 8 Skipped: 0

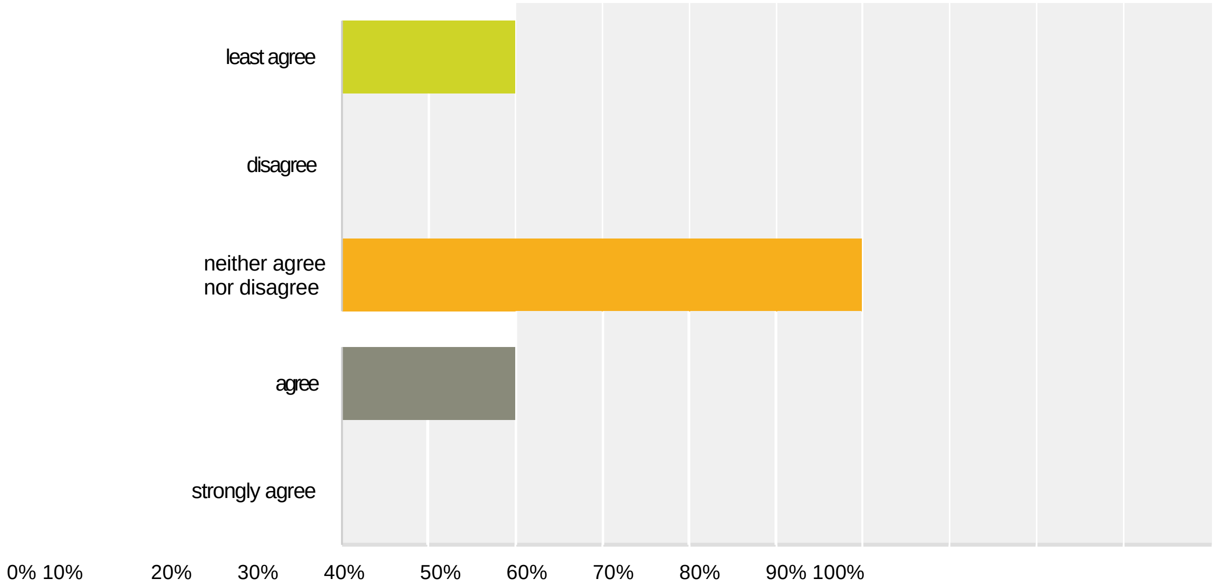


Answer Choices	Responses	
least agree	**0.00%**	0
disagree	**25.00%**	2
neither agree nor disagree	**25.00%**	2
agree	**50.00%**	4
strongly agree	**0.00%**	0
Total Respondents: 8		

#	Please provide rationale	Date
1	Need to know how many patients required to have statistical significance	1/4/2015 11:23 AM
2	Not always possible to provide a power calculation, especially for new diagnostic techniques or for populations who have not been evaluated previously.	1/2/2015 10:43 AM
3	Statistical game sometimes	12/26/2014 5:08 PM

Q2 Radiology Diagnostic Accuracy Trials should explicitly state generalisability of the technique to other vendor equipment for all techniques used.

Answered: 5 Skipped: 3

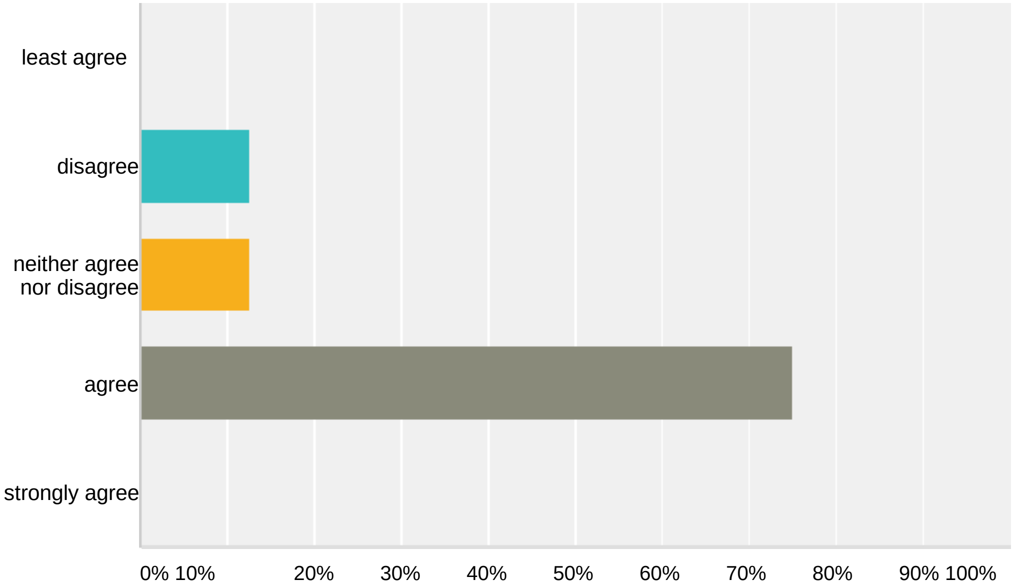


Answer Choices	Responses	
least agree	20.00%	1
disagree	0.00%	0
neither agree nor disagree	60.00%	3
agree	20.00%	1
strongly agree	0.00%	0
Total Respondents: 5		

#	Please provide rationale	Date
1	A study shouldn't be done unless the test can be generalized to all vendor platforms	1/4/2015 11:23 AM
2	Readers should be alerted if the performance of the technique is vendor specific or limited to a given population.	1/2/2015 10:43 AM
3	May be difficult to guess sometimes	12/26/2014 5:08 PM

Q3 Radiology Diagnostic Accuracy Trials should provide robust assessment of inter-observer agreement including >3 observers with variable expertise.

Answered: 8 Skipped: 0



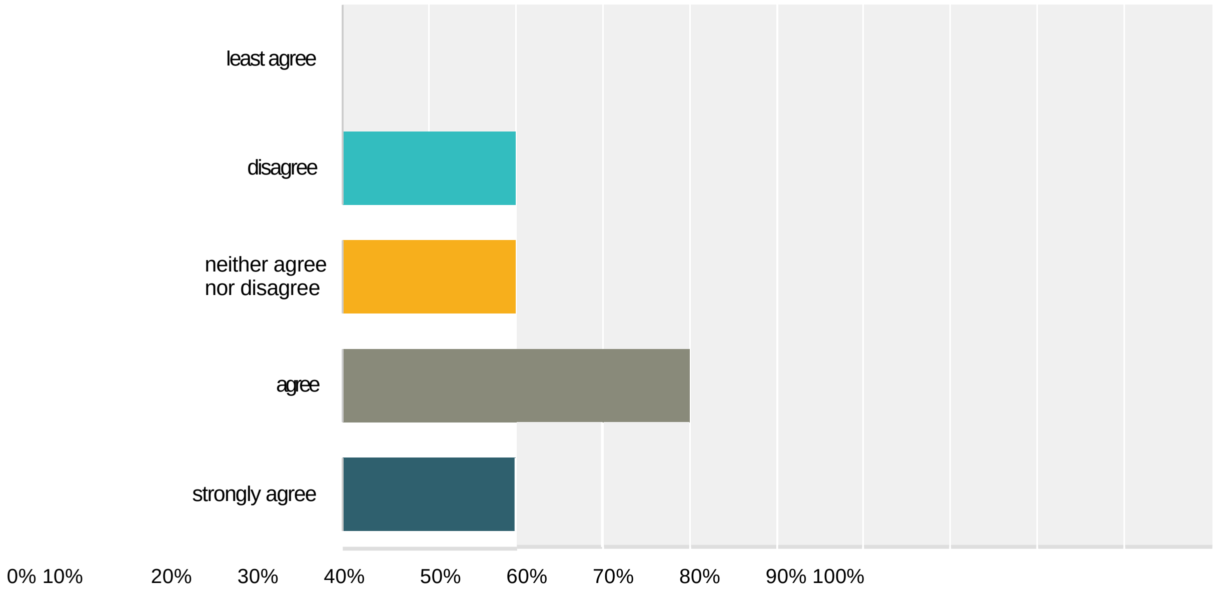
Answer Choices	Responses	
least agree	0.00%	0
disagree	12.50%	1
neither agree nor disagree	12.50%	1
agree	75.00%	6
strongly agree	0.00%	0
Total Respondents: 8		

#	Please provide rationale	Date
2	ideal but sometimes not practical	12/26/2014 5:08 PM

1	Inter-observer agreement is valuable, but as long as the level of expertise is explicitly stated, it should not be mandatory to provide more than two observers.	1/2/2015 10:43 AM

Q4 Changes in the diagnosis after each test should be reported.

Answered: 5 Skipped: 3

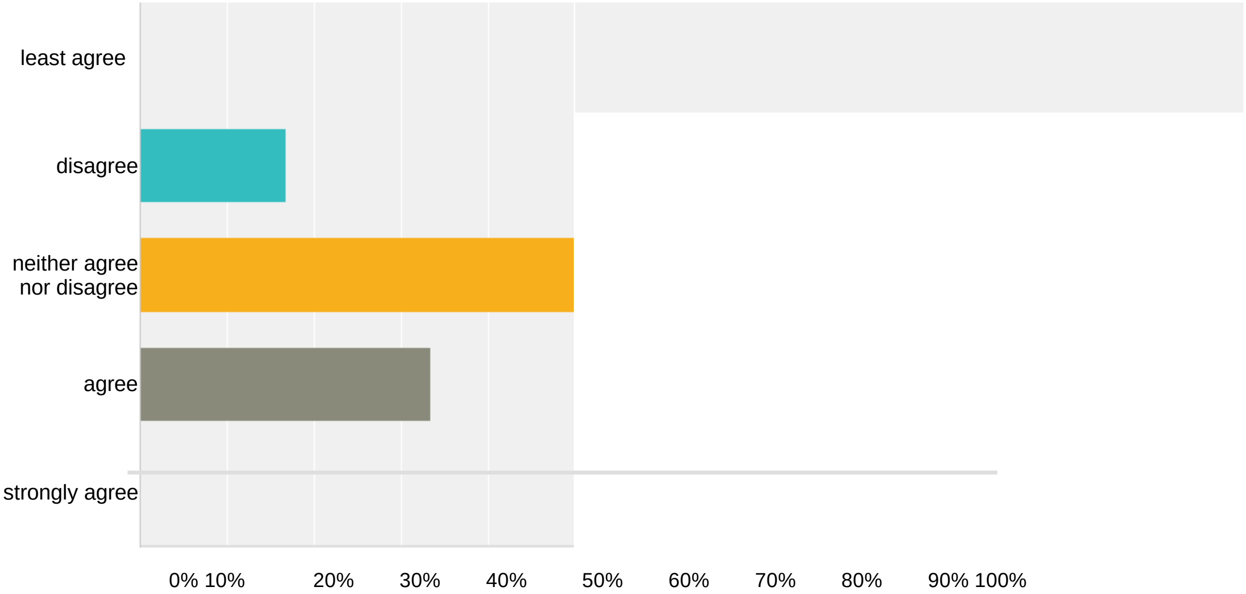


Answer Choices	Responses	
least agree	0.00%	0
disagree	20.00%	1
neither agree nor disagree	20.00%	1
agree	40.00%	2
strongly agree	20.00%	1
Total Respondents: 5		

#	Please provide rationale	Date
	There are no responses.	

Q5 You should know if the level of radiation was less for the index test.

Answered: 6 Skipped: 2

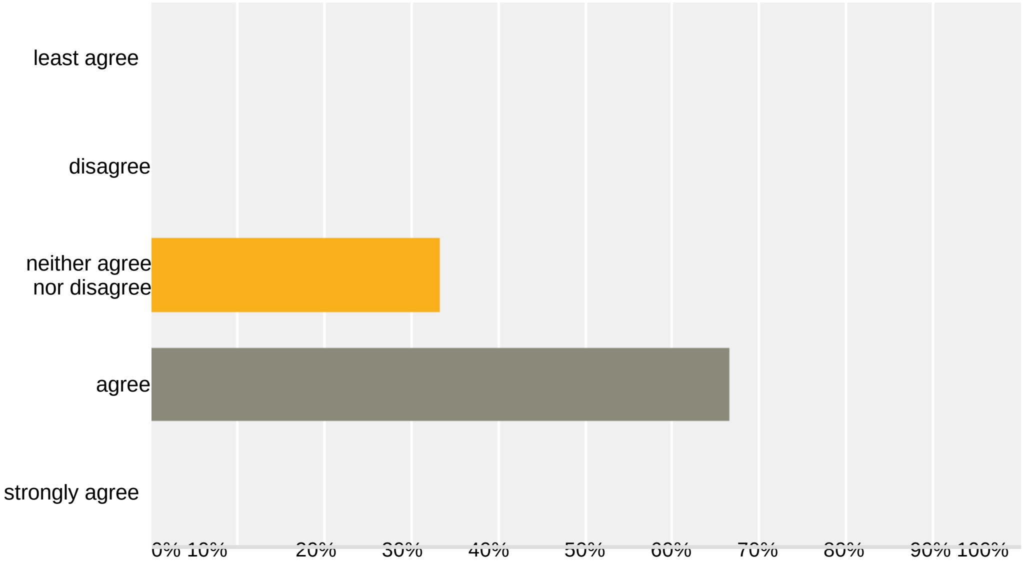


Answer Choices	Responses	
least agree	0.00%	0
disagree	16.67%	1
neither agree nor disagree	50.00%	3
agree	33.33%	2
strongly agree	0.00%	0
Total Respondents: 6		

#	Other (please specify)	Date
1	Should be reported if test involves radiation	1/4/2015 11:23 AM
2	usually not a game changer	12/26/2014 5:08 PM

Q6 Radiology Diagnostic Accuracy Trials should include a cross tabulation of results for the index test and reference standard.

Answered: 6 Skipped: 2

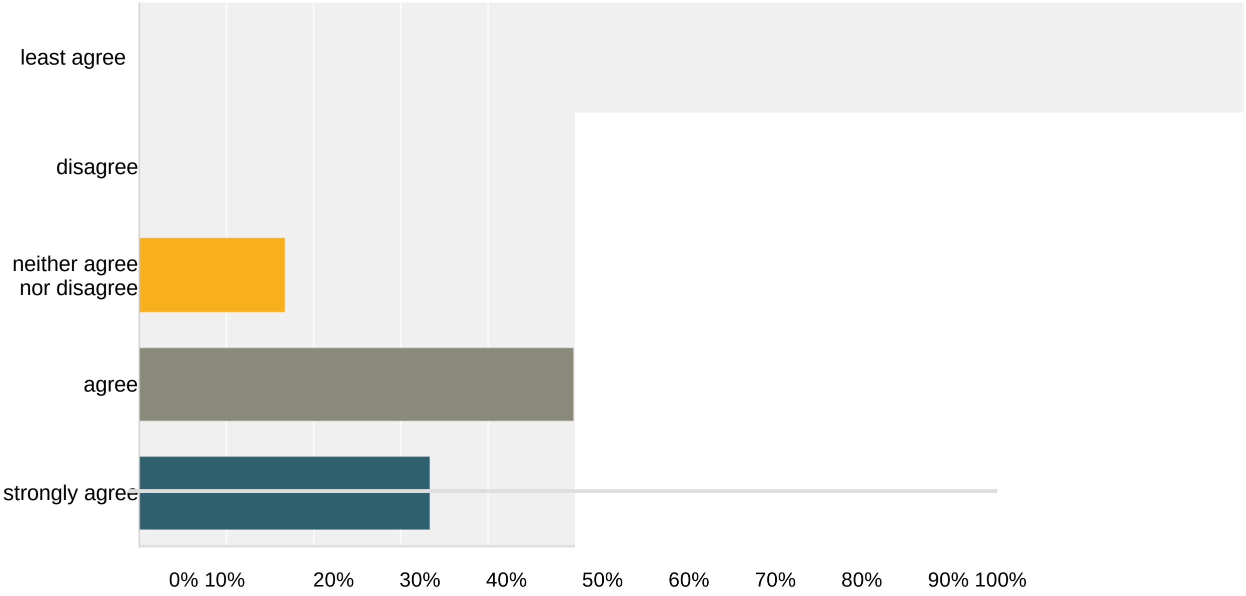


Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	33.33%	2
agree	66.67%	4
strongly agree	0.00%	0
Total Respondents: 6		

#	Other (please specify)	Date
1	I find this question confusing	1/4/2015 11:23 AM

Q7 Radiology Diagnostic Accuracy Trials should include estimates of test reproducibility.

Answered: 6 Skipped: 2



Answer Choices	Responses	
least agree	0.00%	0
disagree	0.00%	0
neither agree nor disagree	16.67%	1
agree	50.00%	3
strongly agree	33.33%	2
Total Respondents: 6		

#	Other (please specify)	Date
	There are no responses.	

Appendix 29: Results of Mann Whitney Analysis

STARD Q 1

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	302.0	288.0	24.230945	16.777778
F	13	194.0	208.0	24.230945	14.923077
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	194.0000
Normal Approximation	
Z	-0.5571
One-Sided Pr < Z	0.2887
Two-Sided Pr > |Z|	0.5774
t Approximation	
One-Sided Pr < Z	0.2908
Two-Sided Pr > |Z|	0.5816
Z includes a continuity correction of 0.5.	

LEGEND

MEAN SCORE: average of the ranks for the two groups. The scores from residents tended to be higher relative to those from Fellows.

Two-Sided PR> |Z|= p-value

STARD Q2

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	303.50	288.0	23.880550	16.861111
F	13	192.50	208.0	23.880550	14.807692
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	192.5000
Normal Approximation	
Z	-0.6281
One-Sided Pr < Z	0.2650
Two-Sided Pr > |Z|	0.5299
t Approximation	
One-Sided Pr < Z	0.2673
Two-Sided Pr > |Z|	0.5347
Z includes a continuity correction of 0.5.	

STARD Q 3

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	266.50	288.0	23.951564	14.805556
F	13	229.50	208.0	23.951564	17.653846
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	229.5000
Normal Approximation	
Z	0.8768
One-Sided Pr > Z	0.1903
Two-Sided Pr > |Z|	0.3806
t Approximation	
One-Sided Pr > Z	0.1938
Two-Sided Pr > |Z|	0.3876
Z includes a continuity correction of 0.5.	

STARD Q 4

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	304.50	288.0	24.311288	16.916667
F	13	191.50	208.0	24.311288	14.730769
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	191.5000
Normal Approximation	
Z	-0.6581
One-Sided Pr < Z	0.2552
Two-Sided Pr > |Z|	0.5105
t Approximation	
One-Sided Pr < Z	0.2577
Two-Sided Pr > |Z|	0.5155
Z includes a continuity correction of 0.5.	

STARD Q 5

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	297.0	279.0	22.516201	16.50
F	12	168.0	186.0	22.516201	14.00

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	168.0000
Normal Approximation	
Z	-0.7772
One-Sided Pr < Z	0.2185
Two-Sided Pr > |Z|	0.4370
t Approximation	
One-Sided Pr < Z	0.2217
Two-Sided Pr > |Z|	0.4433

Z includes a continuity correction of 0.5.

STARD Q6

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	259.0	288.0	24.272446	14.388889
F	13	237.0	208.0	24.272446	18.230769
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	237.000
Normal Approximation	
Z	1.1742
One-Sided Pr > Z	0.1202
Two-Sided Pr > |Z|	0.2403
t Approximation	
One-Sided Pr > Z	0.1248
Two-Sided Pr > |Z|	0.2496
Z includes a continuity correction of 0.5.	

STARD Q 7

Wilcoxon Scores (Rank Sums) for Variable value
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	260.0	288.0	23.975189	14.444444
F	13	236.0	208.0	23.975189	18.153846

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	236.0000
Normal Approximation	
Z	1.1470
One-Sided Pr > Z	0.1257
Two-Sided Pr > |Z|	0.2514
t Approximation	
One-Sided Pr > Z	0.1302
Two-Sided Pr > |Z|	0.2604

Z includes a continuity correction of 0.5.

STARD Q 8

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	17	280.0	263.50	23.189797	16.470588
F	13	185.0	201.50	23.189797	14.230769
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	185.0000
Normal Approximation	
Z	-0.6900
One-Sided Pr < Z	0.2451
Two-Sided Pr > |Z|	0.4902
t Approximation	
One-Sided Pr < Z	0.2479
Two-Sided Pr > |Z|	0.4957
Z includes a continuity correction of 0.5.	

STARD Q9

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	273.50	288.0	23.912138	15.194444
F	13	222.50	208.0	23.912138	17.115385

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	222.5000
Normal Approximation	
Z	0.5855
One-Sided Pr > Z	0.2791
Two-Sided Pr > |Z|	0.5582
t Approximation	
One-Sided Pr > Z	0.2813
Two-Sided Pr > |Z|	0.5626

Z includes a continuity correction of 0.5.

STARD Q10

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	277.0	288.0	24.272446	15.388889
F	13	219.0	208.0	24.272446	16.846154
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	219.000
Normal Approximation	
Z	0.4326
One-Sided Pr > Z	0.3327
Two-Sided Pr > |Z|	0.6653
t Approximation	
One-Sided Pr > Z	0.3342
Two-Sided Pr > |Z|	0.6684
Z includes a continuity correction of 0.5.	

STARD Q11

Wilcoxon Scores (Rank Sums) for Variable value
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	16	228.0	232.0	21.178006	14.250000
F	12	178.0	174.0	21.178006	14.833333

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	178.0000
Normal Approximation	
Z	0.1653
One-Sided Pr > Z	0.4344
Two-Sided Pr > |Z|	0.8687
t Approximation	
One-Sided Pr > Z	0.4350
Two-Sided Pr > |Z|	0.8700

Z includes a continuity correction of 0.5.

STARD Q12

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	271.0	288.0	24.303524	15.055556
F	13	225.0	208.0	24.303524	17.307692
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	225.0000
Normal Approximation	
Z	0.6789
One-Sided Pr > Z	0.2486
Two-Sided Pr > |Z|	0.4972
t Approximation	
One-Sided Pr > Z	0.2512
Two-Sided Pr > |Z|	0.5024
Z includes a continuity correction of 0.5.	

STARD Q13

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	280.0	279.0	22.694675	15.555556
F	12	185.0	186.0	22.694675	15.416667
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	185.0000
Normal Approximation	
Z	-0.0220
One-Sided Pr < Z	0.4912
Two-Sided Pr > |Z|	0.9824
t Approximation	
One-Sided Pr < Z	0.4913
Two-Sided Pr > |Z|	0.9826
Z includes a continuity correction of 0.5.	

STARD Q14

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	275.50	288.0	24.150336	15.305556
F	13	220.50	208.0	24.150336	16.961538

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	220.5000
Normal Approximation	
Z	0.4969
One-Sided Pr > Z	0.3096
Two-Sided Pr > |Z|	0.6193
t Approximation	
One-Sided Pr > Z	0.3114
Two-Sided Pr > |Z|	0.6229

Z includes a continuity correction of 0.5.

STARD Q15

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	263.50	288.0	22.761172	14.638889
F	13	232.50	208.0	22.761172	17.884615
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	232.5000
Normal Approximation	
Z	1.0544
One-Sided Pr > Z	0.1458
Two-Sided Pr > |Z|	0.2917
t Approximation	
One-Sided Pr > Z	0.1501
Two-Sided Pr > |Z|	0.3001
Z includes a continuity correction of 0.5.	

STARD Q16

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	17	257.50	255.0	22.136556	15.147059
F	12	177.50	180.0	22.136556	14.791667
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	177.500
Normal Approximation	
Z	-0.0903
One-Sided Pr < Z	0.4640
Two-Sided Pr > |Z|	0.9280
t Approximation	
One-Sided Pr < Z	0.4643
Two-Sided Pr > |Z|	0.9287
Z includes a continuity correction of 0.5.	

STARD Q17

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	282.50	288.0	24.419717	15.694444
F	13	213.50	208.0	24.419717	16.423077

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	213.5000
Normal Approximation	
Z	0.2048
One-Sided Pr > Z	0.4189
Two-Sided Pr > |Z|	0.8378
t Approximation	
One-Sided Pr > Z	0.4196
Two-Sided Pr > |Z|	0.8391

Z includes a continuity correction of 0.5.

STARD Q18

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	246.0	288.0	24.589140	13.666667
F	13	250.0	208.0	24.589140	19.230769
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	250.0000
Normal Approximation	
Z	1.6877
One-Sided Pr > Z	0.0457
Two-Sided Pr > |Z|	0.0915
t Approximation	
One-Sided Pr > Z	0.0509
Two-Sided Pr > |Z|	0.1018
Z includes a continuity correction of 0.5.	

STARD Q19

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	268.0	288.0	24.463468	14.888889
F	13	228.0	208.0	24.463468	17.538462

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	228.0000
Normal Approximation	
Z	0.7971
One-Sided Pr > Z	0.2127
Two-Sided Pr > |Z|	0.4254
t Approximation	
One-Sided Pr > Z	0.2158
Two-Sided Pr > |Z|	0.4317

Z includes a continuity correction of 0.5.

STARD Q20

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	282.0	279.0	23.171028	15.666667
F	12	183.0	186.0	23.171028	15.250000
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	183.0000
Normal Approximation	
Z	-0.1079
One-Sided Pr < Z	0.4570
Two-Sided Pr > |Z|	0.9141
t Approximation	
One-Sided Pr < Z	0.4574
Two-Sided Pr > |Z|	0.9148
Z includes a continuity correction of 0.5.	

STARD Q21

Wilcoxon Scores (Rank Sums) for Variable value
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	285.0	288.0	23.674200	15.833333
F	13	211.0	208.0	23.674200	16.230769

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	211.0000
Normal Approximation	
Z	0.1056
One-Sided Pr > Z	0.4579
Two-Sided Pr > |Z|	0.9159
t Approximation	
One-Sided Pr > Z	0.4583
Two-Sided Pr > |Z|	0.9166

Z includes a continuity correction of 0.5.

STARD Q22

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	294.0	288.0	24.548175	16.333333
F	13	202.0	208.0	24.548175	15.538462
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	202.0000
Normal Approximation	
Z	-0.2240
One-Sided Pr < Z	0.4114
Two-Sided Pr > |Z|	0.8227
t Approximation	
One-Sided Pr < Z	0.4121
Two-Sided Pr > |Z|	0.8242
Z includes a continuity correction of 0.5.	

STARD Q23

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group					
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	288.0	288.0	24.619819	16.0
F	13	208.0	208.0	24.619819	16.0
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	208.0000
Normal Approximation	
Z	0.0000
One-Sided Pr < Z	0.5000
Two-Sided Pr > |Z|	1.0000
t Approximation	
One-Sided Pr < Z	0.5000
Two-Sided Pr > |Z|	1.0000
Z includes a continuity correction of 0.5.	

STARD Q24

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	303.50	279.0	22.389037	16.861111
F	12	161.50	186.0	22.389037	13.458333

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	161.5000
Normal Approximation	
Z	-1.0720
One-Sided Pr < Z	0.1419
Two-Sided Pr > |Z|	0.2837
t Approximation	
One-Sided Pr < Z	0.1463
Two-Sided Pr > |Z|	0.2926

Z includes a continuity correction of 0.5.

STARD Q25

Wilcoxon Scores (Rank Sums) for Variable value Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
R	18	288.0	288.0	22.959291	16.0
F	13	208.0	208.0	22.959291	16.0

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic	208.0000
Normal Approximation	
Z	0.0000
One-Sided Pr < Z	0.5000
Two-Sided Pr > |Z|	1.0000
t Approximation	
One-Sided Pr < Z	0.5000
Two-Sided Pr > |Z|	1.0000

Z includes a continuity correction of 0.5.

Appendix 30: Results of Chi- Square Analysis

Table Analysis Q1

Frequency
Percent Row
Pct Col Pct

LEGEND

25 tables, 1 per page, for the 25 questions.
The r = 0 row is the row for Fellows responses;
the r = 1 row is for residents.
The c=0 row is the count of scores less than 7 (0 to 6);
the c=1 row is the count
of responses of 7 or more.
13 of the 15 fellows who responded gave a score of 7 or higher;
16 of the 18 residents who responded gave a score of 7 or higher.
Same format for all 25 tables.

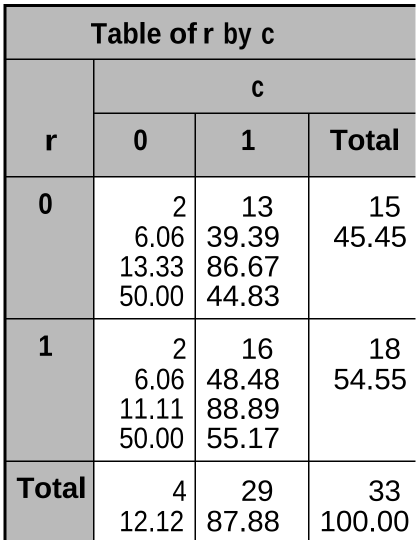
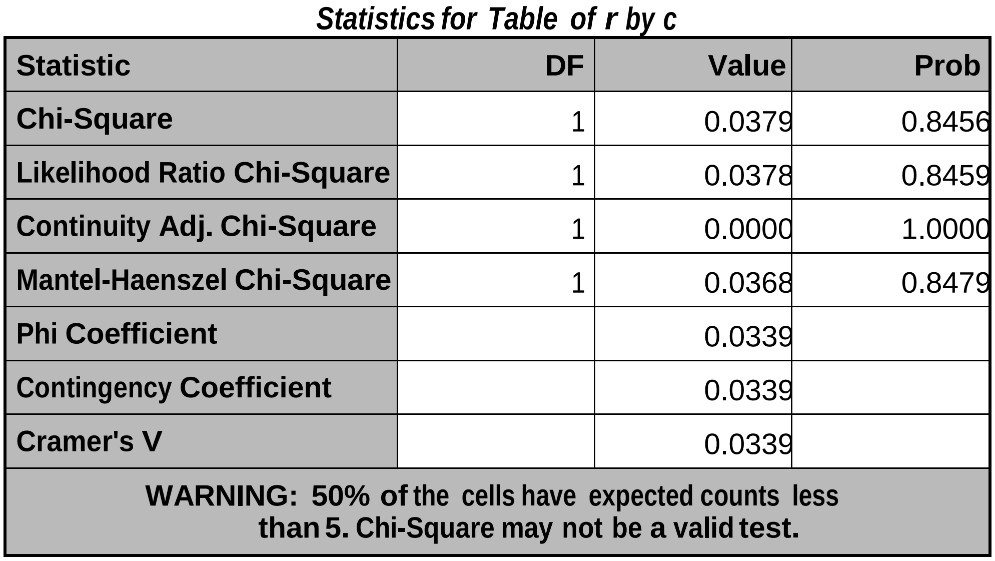


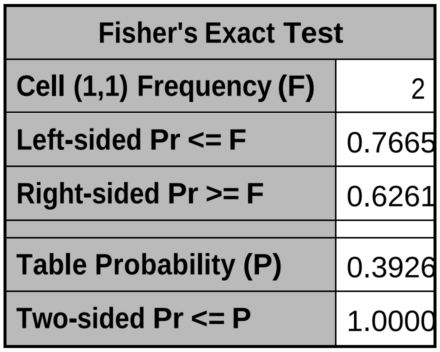
Table of r by c			
	c		
r	0	1	Total
0	2 6.06 13.33 50.00	13 39.39 86.67 44.83	15 45.45
1	2 6.06 11.11 50.00	16 48.48 88.89 55.17	18 54.55
Total	4 12.12	29 87.88	33 100.00



Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0379	0.8456
Likelihood Ratio Chi-Square	1	0.0378	0.8459
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0368	0.8479
Phi Coefficient		0.0339	
Contingency Coefficient		0.0339	
Cramer's V		0.0339	

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.



Fisher's Exact Test	
Cell (1,1) Frequency (F)	2
Left-sided Pr <= F	0.7665
Right-sided Pr >= F	0.6261
Table Probability (P)	0.3926
Two-sided Pr <= P	1.0000

Table Analysis Q2

Table of r by c			
	c		
r	0	1	Total
0	2 6.06 13.33 66.67	13 39.39 86.67 43.33	15 45.45
1	1 3.03 5.56 33.33	17 51.52 94.44 56.67	18 54.55
Total	3 9.09	30 90.91	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.5989	0.4390
Likelihood Ratio Chi-Square	1	0.6016	0.4380
Continuity Adj. Chi-Square	1	0.0275	0.8683
Mantel-Haenszel Chi-Square	1	0.5807	0.4460
Phi Coefficient		0.1347	
Contingency Coefficient		0.1335	
Cramer's V		0.1347	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2
Left-sided Pr <= F	0.9166
Right-sided Pr >= F	0.4298
Table Probability (P)	0.3464
Two-sided Pr <= P	0.5794

Table Analysis Q3

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 75.00	12 36.36 80.00 41.38	15 45.45
1	1 3.03 5.56 25.00	17 51.52 94.44 58.62	18 54.55
Total	4 12.12	29 87.88	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	1.6026	0.2055
Likelihood Ratio Chi-Square	1	1.6398	0.2004
Continuity Adj. Chi-Square	1	0.5334	0.4652
Mantel-Haenszel Chi-Square	1	1.5540	0.2125
Phi Coefficient		0.2204	
Contingency Coefficient		0.2152	
Cramer's V		0.2204	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.9666
Right-sided Pr >= F	0.2335
Table Probability (P)	0.2001
Two-sided Pr <= P	0.3083

Table Analysis Q4

Table of r by c			
	c		
r	0	1	Total
0	3 9.38 21.43 60.00	11 34.38 78.57 40.74	14 43.75
1	2 6.25 11.11 40.00	16 50.00 88.89 59.26	18 56.25
Total	5 15.63	27 84.38	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.6359	0.4252
Likelihood Ratio Chi-Square	1	0.6313	0.4269
Continuity Adj. Chi-Square	1	0.0941	0.7591
Mantel-Haenszel Chi-Square	1	0.6160	0.4325
Phi Coefficient		0.1410	
Contingency Coefficient		0.1396	
Cramer's V		0.1410	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.9006
Right-sided Pr >= F	0.3760
Table Probability (P)	0.2766
Two-sided Pr <= P	0.6313

Table Analysis Q5

Table of r by c			
	c		
r	0	1	Total
0	4 12.50 28.57 80.00	10 31.25 71.43 37.04	14 43.75
1	1 3.13 5.56 20.00	17 53.13 94.44 62.96	18 56.25
Total	5 15.63	27 84.38	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	3.1643	0.0753
Likelihood Ratio Chi-Square	1	3.2618	0.0709
Continuity Adj. Chi-Square	1	1.6593	0.1977
Mantel-Haenszel Chi-Square	1	3.0654	0.0800
Phi Coefficient		0.3145	
Contingency Coefficient		0.3000	
Cramer's V		0.3145	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.9901
Right-sided Pr >= F	0.0994
Table Probability (P)	0.0895
Two-sided Pr <= P	0.1420

Table Analysis Q6

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 42.86	12 36.36 80.00 46.15	15 45.45
1	4 12.12 22.22 57.14	14 42.42 77.78 53.85	18 54.55

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0242	0.8764
Likelihood Ratio Chi-Square	1	0.0242	0.8763
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0234	0.8783
Phi Coefficient		-0.0271	
Contingency Coefficient		0.0271	
Cramer's V		-0.0271	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.6091
Right-sided Pr >= F	0.7168
Table Probability (P)	0.3259
Two-sided Pr <= P	1.0000

Table Analysis Q7

Table of r by c			
	c		
r	0	1	Total
0	1 3.03 6.67 16.67	14 42.42 93.33 51.85	15 45.45
1	5 15.15 27.78 83.33	13 39.39 72.22 48.15	18 54.55
Total	6 18.18	27 81.82	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	2.4512	0.1174
Likelihood Ratio Chi-Square	1	2.6750	0.1019
Continuity Adj. Chi-Square	1	1.2375	0.2660
Mantel-Haenszel Chi-Square	1	2.3770	0.1231
Phi Coefficient		-0.2725	
Contingency Coefficient		0.2630	
Cramer's V		-0.2725	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.1328
Right-sided Pr >= F	0.9832
Table Probability (P)	0.1160
Two-sided Pr <= P	0.1861

Table Analysis Q8

Table of r by c

r	c		
	0	1	Total
0	3 9.38 20.00 75.00	12 37.50 80.00 42.86	15 46.88
1	1 3.13 5.88 25.00	16 50.00 94.12 57.14	17 53.13
Total	4 12.50	28 87.50	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	1.4521	0.2282
Likelihood Ratio Chi-Square	1	1.4948	0.2215
Continuity Adj. Chi-Square	1	0.4482	0.5032
Mantel-Haenszel Chi-Square	1	1.4067	0.2356
Phi Coefficient		0.2130	
Contingency Coefficient		0.2083	
Cramer's V		0.2130	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.9620
Right-sided Pr >= F	0.2531
Table Probability (P)	0.2151
Two-sided Pr <= P	0.3192

Table Analysis Q9

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 50.00	12 36.36 80.00 44.44	15 45.45
1	3 9.09 16.67 50.00	15 45.45 83.33 55.56	18 54.55
Total	6 18.18	27 81.82	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0611	0.8047
Likelihood Ratio Chi-Square	1	0.0609	0.8051
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0593	0.8077
Phi Coefficient		0.0430	
Contingency Coefficient		0.0430	
Cramer's V		0.0430	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.7581
Right-sided Pr >= F	0.5771
Table Probability (P)	0.3352
Two-sided Pr <= P	1.0000

Table Analysis Q10

Table of r by c			
	c		
r	0	1	Total
0	2 6.06 13.33 28.57	13 39.39 86.67 50.00	15 45.45
1	5 15.15 27.78 71.43	13 39.39 72.22 50.00	18 54.55
Total	7 21.21	26 78.79	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	1.0214	0.3122
Likelihood Ratio Chi-Square	1	1.0552	0.3043
Continuity Adj. Chi-Square	1	0.3400	0.5598
Mantel-Haenszel Chi-Square	1	0.9905	0.3196
Phi Coefficient		-0.1759	
Contingency Coefficient		0.1733	
Cramer's V		-0.1759	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2
Left-sided Pr <= F	0.2832
Right-sided Pr >= F	0.9274
Table Probability (P)	0.2106
Two-sided Pr <= P	0.4134

Table Analysis Q11

Table of r by c			
	c		
r	0	1	Total
0	5 16.67 35.71 45.45	9 30.00 64.29 47.37	14 46.67
1	6 20.00 37.50 54.55	10 33.33 62.50 52.63	16 53.33
Total	11 36.67	19 63.33	30 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0103	0.9193
Likelihood Ratio Chi-Square	1	0.0103	0.9193
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0099	0.9207
Phi Coefficient		-0.0185	
Contingency Coefficient		0.0185	
Cramer's V		-0.0185	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	5
Left-sided Pr <= F	0.6101
Right-sided Pr >= F	0.6834
Table Probability (P)	0.2935
Two-sided Pr <= P	1.0000

Table Analysis Q12

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 50.00	12 36.36 80.00 44.44	15 45.45
1	3 9.09 16.67 50.00	15 45.45 83.33 55.56	18 54.55
Total	6 18.18	27 81.82	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0611	0.8047
Likelihood Ratio Chi-Square	1	0.0609	0.8051
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0593	0.8077
Phi Coefficient		0.0430	
Contingency Coefficient		0.0430	
Cramer's V		0.0430	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.7581
Right-sided Pr >= F	0.5771
Table Probability (P)	0.3352
Two-sided Pr <= P	1.0000

Table Analysis Q13

Table of r by c

	c		
r	0	1	Total
0	2 6.25 14.29 50.00	12 37.50 85.71 42.86	14 43.75
1	2 6.25 11.11 50.00	16 50.00 88.89 57.14	18 56.25
Total	4 12.50	28 87.50	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0726	0.7876
Likelihood Ratio Chi-Square	1	0.0721	0.7883
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0703	0.7909
Phi Coefficient		0.0476	
Contingency Coefficient		0.0476	
Cramer's V		0.0476	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2
Left-sided Pr <= F	0.7900
Right-sided Pr >= F	0.5972
Table Probability (P)	0.3872
Two-sided Pr <= P	1.0000

Table Analysis Q14

Table of r by c			
	c		
r	0	1	Total
0	1 3.03 6.67 33.33	14 42.42 93.33 46.67	15 45.45
1	2 6.06 11.11 66.67	16 48.48 88.89 53.33	18 54.55
Total	3 9.09	30 90.91	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.1956	0.6583
Likelihood Ratio Chi-Square	1	0.2001	0.6546
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.1896	0.6632
Phi Coefficient		-0.0770	
Contingency Coefficient		0.0768	
Cramer's V		-0.0770	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.5702
Right-sided Pr >= F	0.8504
Table Probability (P)	0.4206
Two-sided Pr <= P	1.0000

Table Analysis Q15

Table of r by c			
	c		
r	0	1	Total
0	1 3.03 6.67 20.00	14 42.42 93.33 50.00	15 45.45
1	4 12.12 22.22 80.00	14 42.42 77.78 50.00	18 54.55
Total	5 15.15	28 84.85	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	1.5400	0.2146
Likelihood Ratio Chi-Square	1	1.6543	0.1984
Continuity Adj. Chi-Square	1	0.5677	0.4512
Mantel-Haenszel Chi-Square	1	1.4933	0.2217
Phi Coefficient		-0.2160	
Contingency Coefficient		0.2112	
Cramer's V		-0.2160	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.2295
Right-sided Pr >= F	0.9639
Table Probability (P)	0.1934
Two-sided Pr <= P	0.3457

Table Analysis Q16

Table of r by c			
	c		
r	0	1	Total
0	4 12.90 28.57 50.00	10 32.26 71.43 43.48	14 45.16
1	4 12.90 23.53 50.00	13 41.94 76.47 56.52	17 54.84
Total	8 25.81	23 74.19	31 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.1019	0.7495
Likelihood Ratio Chi-Square	1	0.1016	0.7499
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0986	0.7535
Phi Coefficient		0.0573	
Contingency Coefficient		0.0572	
Cramer's V		0.0573	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.7679
Right-sided Pr >= F	0.5341
Table Probability (P)	0.3020
Two-sided Pr <= P	1.0000

Table Analysis Q17

Table of r by c			
	c		
r	0	1	Total
0	5 15.15 33.33 62.50	10 30.30 66.67 40.00	15 45.45
1	3 9.09 16.67 37.50	15 45.45 83.33 60.00	18 54.55
Total	8 24.24	25 75.76	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	1.2375	0.2660
Likelihood Ratio Chi-Square	1	1.2390	0.2657
Continuity Adj. Chi-Square	1	0.4964	0.4811
Mantel-Haenszel Chi-Square	1	1.2000	0.2733
Phi Coefficient		0.1936	
Contingency Coefficient		0.1901	
Cramer's V		0.1936	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	5
Left-sided Pr <= F	0.9360
Right-sided Pr >= F	0.2405
Table Probability (P)	0.1765
Two-sided Pr <= P	0.4184

Table Analysis Q18

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 42.86	12 36.36 80.00 46.15	15 45.45
1	4 12.12 22.22 57.14	14 42.42 77.78 53.85	18 54.55
Total	7 21.21	26 78.79	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0242	0.8764
Likelihood Ratio Chi-Square	1	0.0242	0.8763
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0234	0.8783
Phi Coefficient		-0.0271	
Contingency Coefficient		0.0271	
Cramer's V		-0.0271	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.6091
Right-sided Pr >= F	0.7168
Table Probability (P)	0.3259
Two-sided Pr <= P	1.0000

Table Analysis Q19

Table of r by c			
	c		
r	0	1	Total
0	3 9.09 20.00 37.50	12 36.36 80.00 48.00	15 45.45
1	5 15.15 27.78 62.50	13 39.39 72.22 52.00	18 54.55
Total	8 24.24	25 75.76	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.2695	0.6037
Likelihood Ratio Chi-Square	1	0.2722	0.6018
Continuity Adj. Chi-Square	1	0.0124	0.9114
Mantel-Haenszel Chi-Square	1	0.2613	0.6092
Phi Coefficient		-0.0904	
Contingency Coefficient		0.0900	
Cramer's V		-0.0904	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	3
Left-sided Pr <= F	0.4587
Right-sided Pr >= F	0.8221
Table Probability (P)	0.2808
Two-sided Pr <= P	0.6992

Table Analysis 20

Table of r by c			
	c		
r	0	1	Total
0	5 15.63 35.71 55.56	9 28.13 64.29 39.13	14 43.75
1	4 12.50 22.22 44.44	14 43.75 77.78 60.87	18 56.25
Total	9 28.13	23 71.88	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.7091	0.3997
Likelihood Ratio Chi-Square	1	0.7057	0.4009
Continuity Adj. Chi-Square	1	0.1988	0.6557
Mantel-Haenszel Chi-Square	1	0.6870	0.4072
Phi Coefficient		0.1489	
Contingency Coefficient		0.1472	
Cramer's V		0.1489	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	5
Left-sided Pr <= F	0.8919
Right-sided Pr >= F	0.3265
Table Probability (P)	0.2184
Two-sided Pr <= P	0.4533

Table Analysis 21

Table of r by c			
	c		
r	0	1	Total
0	1 3.03 6.67 50.00	14 42.42 93.33 45.16	15 45.45
1	1 3.03 5.56 50.00	17 51.52 94.44 54.84	18 54.55
Total	2 6.06	31 93.94	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0177	0.8940
Likelihood Ratio Chi-Square	1	0.0177	0.8942
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0172	0.8956
Phi Coefficient		0.0232	
Contingency Coefficient		0.0232	
Cramer's V		0.0232	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.8011
Right-sided Pr >= F	0.7102
Table Probability (P)	0.5114
Two-sided Pr <= P	1.0000

Table Analysis 22

Table of r by c			
	c		
r	0	1	Total
0	4 12.12 26.67 50.00	11 33.33 73.33 44.00	15 45.45
1	4 12.12 22.22 50.00	14 42.42 77.78 56.00	18 54.55
Total	8 24.24	25 75.76	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0880	0.7667
Likelihood Ratio Chi-Square	1	0.0878	0.7670
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0853	0.7702
Phi Coefficient		0.0516	
Contingency Coefficient		0.0516	
Cramer's V		0.0516	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.7595
Right-sided Pr >= F	0.5413
Table Probability (P)	0.3008
Two-sided Pr <= P	1.0000

Table Analysis 23

Table of r by c			
	c		
r	0	1	Total
0	5 15.15 33.33 41.67	10 30.30 66.67 47.62	15 45.45
1	7 21.21 38.89 58.33	11 33.33 61.11 52.38	18 54.55
Total	12 36.36	21 63.64	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.1091	0.7411
Likelihood Ratio Chi-Square	1	0.1094	0.7408
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.1058	0.7450
Phi Coefficient		-0.0575	
Contingency Coefficient		0.0574	
Cramer's V		-0.0575	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	5
Left-sided Pr <= F	0.5144
Right-sided Pr >= F	0.7550
Table Probability (P)	0.2693
Two-sided Pr <= P	1.0000

Table Analysis 24

Table of r by c			
	c		
r	0	1	Total
0	1 3.13 7.14 50.00	13 40.63 92.86 43.33	14 43.75
1	1 3.13 5.56 50.00	17 53.13 94.44 56.67	18 56.25
Total	2 6.25	30 93.75	32 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	0.0339	0.8540
Likelihood Ratio Chi-Square	1	0.0336	0.8545
Continuity Adj. Chi-Square	1	0.0000	1.0000
Mantel-Haenszel Chi-Square	1	0.0328	0.8563
Phi Coefficient		0.0325	
Contingency Coefficient		0.0325	
Cramer's V		0.0325	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.8165
Right-sided Pr >= F	0.6915
Table Probability (P)	0.5081
Two-sided Pr <= P	1.0000

Table Analysis Q25

Table of r by c			
	c		
r	0	1	Total
0	2 6.06 13.33 100.00	13 39.39 86.67 41.94	15 45.45
1	0 0.00 0.00 0.00	18 54.55 100.00 58.06	18 54.55
Total	2 6.06	31 93.94	33 100.00

Statistics for Table of r by c

Statistic	DF	Value	Prob
Chi-Square	1	2.5548	0.1100
Likelihood Ratio Chi-Square	1	3.3095	0.0689
Continuity Adj. Chi-Square	1	0.7496	0.3866
Mantel-Haenszel Chi-Square	1	2.4774	0.1155
Phi Coefficient		0.2782	
Contingency Coefficient		0.2681	
Cramer's V		0.2782	
WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	0.1989
Table Probability (P)	0.1989
Two-sided Pr <= P	0.1989