Investigating the attainment of optimum data quality for EHR Big Data: proposing a new methodological approach

Suraj Juddoo
M00389225

School of Science and Technology
Middlesex University

A thesis submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Director of Studies: Dr Carlisle George

Supervisors: Dr David Windridge & Prof Miltos Petridis

January 2022

Acknowledgement

This research started with my fascination with how a very simple concept of data use has evolved to become a highly complex area. This fascination would have stayed just as an area of interest if there were not some key persons. These persons gave me the necessary drive during the lonely and uncertain periods typical of doctoral research. Over and above the drive and encouragements, these persons also brought about their research and technical expertise which calibrated and fine-tuned the progress of the research study.

Dr Carlisle George is the first person who deserves my deepest gratitude. He has been alongside the progress of this research all the way, and it was a long way. He has this capacity to encourage me with his calm but very methodical approach. He has the knack to find the right words at the right time, even if it was distance supervision. I always knew I had his support. His objective insights into how to approach this research study provided me the necessary confidence to know that the study will not be negatively affected by my own bias.

Along Dr George, there were also very important supervisory members who helped me progress through the different phases of the research study. **Dr Penny Duquenoy** encouraged me and framed the initial research framework. The decision to investigate deeper towards data governance should be credited to her. **Dr David Windridge** brought his high expertise mostly towards the use of machine learning in the study. Late **Professor Miltos Petridis** was also very helpful towards discussing the implications of involving machine learning for data quality in general. Without a doubt, they all have an immense contribution towards shaping this research study.

Family support is essential for such a long-term endeavour. My wife Poonam provided me her wisdom during the trying times. She has such a pragmatic approach towards life which is simply inspiring. My parents for always enquiring and giving general encouragements towards going forward.

Abstract

The value derivable from the use of data is continuously increasing since some years. Both commercial and non-commercial organisations have realised the immense benefits that might be derived if all data at their disposal could be analysed and form the basis of decision taking. The technological tools required to produce, capture, store, transmit and analyse huge amounts of data form the background to the development of the phenomenon of Big Data. With Big Data, the aim is to be able to generate value from huge amounts of data, often in non-structured format and produced extremely frequently. However, the potential value derivable depends on general level of governance of data, more precisely on the quality of the data. The field of data quality is well researched for traditional data uses but is still in its infancy for the Big Data context. This dissertation focused on investigating effective methods to enhance data quality for Big Data. The principal deliverable of this research is in the form of a methodological approach which can be used to optimize the level of data quality in the Big Data context. Since data quality is contextual, (that is a non-generalizable field), this research study focuses on applying the methodological approach in one use case, in terms of the Electronic Health Records (EHR).

The first main contribution to knowledge of this study systematically investigates which data quality dimensions (DQDs) are most important for EHR Big Data. The two most important dimensions ascertained by the research methods applied in this study are **accuracy** and **completeness.** These are two well-known dimensions, and this study confirms that they are also very important for EHR Big Data. The second important contribution to knowledge is an investigation into whether Artificial Intelligence with a special focus upon machine learning could be used in improving the detection of dirty data, focusing on the two data quality dimensions of accuracy and completeness. Regression and clustering algorithms proved to be more adequate for accuracy and completeness related issues respectively, based on the experiments carried out. However, the limits of implementing and using machine learning algorithms for detecting data quality issues for Big Data were also revealed and discussed in this research study. It can safely be deduced from the knowledge derived from this part of the research study that use of machine learning for enhancing data quality issues detection is a promising area but not yet a panacea which automates this entire process. The third important contribution is a proposed guideline to undertake data repairs most efficiently for Big Data; this involved surveying and comparing existing data cleansing algorithms against a prototype developed for data reparation. Weaknesses of existing algorithms are highlighted and are considered as areas of practice which efficient data reparation algorithms must focus upon.

Those three important contributions form the nucleus for a new data quality methodological approach which could be used to optimize Big Data quality, as applied in the context of EHR. Some of the activities and techniques discussed through the proposed methodological approach can be transposed to other industries and use cases to a large extent. The proposed data quality methodological approach can be used by practitioners of Big Data Quality who follow a data-driven strategy. As opposed to existing Big Data quality frameworks, the proposed data quality methodological approach has the advantage of being more precise and specific. It gives clear and proven methods to undertake the main identified stages of a Big Data quality lifecycle and therefore can be applied by practitioners in the area.

This research study provides some promising results and deliverables. It also paves the way for further research in the area. Technical and technological changes in Big Data is rapidly evolving and future research should be focusing on new representations of Big Data, the real-time streaming aspect, and replicating same research methods used in this current research study but on new technologies to validate current results.

Contents

List of abbreviations

ACM	Association for Computing Machinery
AHRF	Area Health Resource Files
AI	Artificial Intelligence
AIMQ	A methodology for information quality assessment
APC	Ambulatory Payment Classification
API	Application Programming Interface
AUC	Area Under the Curve
BBC	British Broadcasting Corporation
BDQMA	Big Data Quality Methodological Approach
BDQPF	Big Data quality framework
BRFSS	Behavioral Risk Factor Surveillance System
BSON	Binary JSON
CAGR	Compound Annual Growth Rate
CAIRAD	Co-appearance based Analysis for Incorrect Records and Attribute-values Detection
CART	Classification and Regression Trees
CBMS	Cluster-Based Best Match Scanning
CDC	Centre for Disease Control and Prevention
CFD	Conditional Functional Dependency
CIHI	Canadian Institute for Health Information
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality)
CPT	Current Procedural Terminology
CRI	Clustering based Random Imputation
CSV	comma-separated values
DA	Detection Accuracy
DBMS	Database Management Systems
DC	Denial Constraints
DO	Doctor of Osteopathic Medicine
DQ	Data Quality
DQD	Data Quality Dimension
DQF4CT	Data Quality issues in Classification Tasks
DQR	Data quality rules
EBM	Evidence Based Medicine
HER	Electronic Health Records
FD	Functional dependencies
Febrl	Freely extensible biomedical record linkage
FPR	False Positive Rate
GAIN	Generative Adversarial Imputation Nets
GAN	Generative Adversarial Nets
GBDT	Gradient Boosting Decision Tree
GFS	Google File System

GIGO	Garbage In Garbage Out
GUI	Graphical User Interface
HCPCS	Healthcare Common Procedure Coding System
HDFS	Hadoop Distribute File System
HPI	Health Provider Index
HTML	Hypertext Markup Language
IBM	International Business Machines Corporation
IC	Integrated Constraints
ICD	International Classification of Diseases
ID	Identity Number
IHC	Inner Hermeneutic Cycle
IoT	Internet of Things
IQ	Information Quality
IS	Information Systems
ITS	Intelligent Transportation Systems
JDIQ	Journal of Data and Information Quality
JSON	JavaScript Object Notation
KNN	K-Nearest Neighbours
LLS	Local Least Squares
LOF	Local Outlier Factor
LOINC	Logical Observation Identifiers Names and Codes
LRMC	Low Rank Matrix Completion
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
M	Million
MD	Doctor of Medicine
MDC	Major Diagnosis Category
MEDLINE	Medical Literature Analysis and Retrieval System Online
ML	Machine Learning
NDC	National Drug Code
NPI	National Provider Index
NZHIS	New Zealand Health Information Service
ODBC	Open Database Connectivity
PDF	Portable Document Format
PDSS	Personalized Decision Support System
POSMAD	planning, obtaining, storing and sharing, maintaining, applying, and disposing of data
PPCA	Probabilistic Principal Component Analysis
RAM	Random Access Memory
RDF	Resource Description Framework
RMSE	Root Mean Squared Error
RO	Research Objective
ROC	Receiver Operating Characteristics

RQ	Research Question
SNAP	Smoking, nutrition, alcohol, physical activity
SQL	Structured Query Language
SRS-l_p	l_p-norm regularised sparse self-representation
SVD	Single Value Decomposition
SVM	Support Vector Machine
TB	Terabytes
TDQM	Total Data Quality Management
TF-IDF	Term Frequency–Inverse Document Frequency
TIQM	Total Information Quality Management
TPR	True Positive Rate
UDFs	User Defined Functions
UNECE	United Nations Economic Commission for Europe
XML	Extensible Markup Language
ZB	Zetabytes

List of tables

List of figures

List of research publications

The following is a list of research outputs presented at various conferences and journal articles which relate to this doctoral study:

- Juddoo, S., and George, C.,2020. A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy issues to improve Data Analytics in Big Data for the Healthcare Industry. *Proceedings of the 3rd IEEE International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM 2020)*. 25-27 November 2020, University of Mauritius.
- Juddoo, S., and George, C.,2018. Discovering most important Data Quality Dimensions using Latent Semantic Analysis, *Conference Proceedings of International Conference on advances in Big Data, Computing and Data Communication Systems (icABCD)*, IEEExplore
- Juddoo, S., George, C., Duquenoy, P., and Windridge, D.,2018. Data Governance in health industry: Investigating data quality dimensions within a Big Data context, *Applied System Innovation*, *1(4)*, 43; https://doi.org/10.3390/asi1040043
- Juddoo, S., George, C., Duquenoy, P., and Windridge, D.,2017. Data Governance in Healthcare: Investigating data quality dimensions within a big data context, *TILTing perspectives 2017*
- Juddoo, S.,2015. Overview of Big Data quality challenges, *3rd ERPBSS international conference.*
- Juddoo, S.,2015. Overview of Big Data quality challenges, *ICCCS IEEExplore conference proceedings* , doi: 10.1109/CCCS.2015.7374131

Chapter 1: Introduction

1.1 Introduction

This chapter aims to set the scene for the research presented in this thesis. The background area of the work undertaken is discussed in terms of two main areas, namely data quality and Big Data. The importance of Big Data and its general characteristics is described along with the general description of data use in the health industry. Furthermore, the exact nature of the area of the research, that is, the development of a data quality methodological approach to improve data quality for EHR Big Data will be discussed. Since enforcing data quality is an arduous and very contextual undertaking in many cases, having a clearly understood and systematic data driven approach is considered beneficial for practitioners to reduce its cost. One of the main knowledge gaps that this research attempts to tackle is investigating the possibility of an optimised methodological approach to improve data quality for Big Data. The research domain of data quality for Big Data is a novel one, with many questions still pending regarding how best to achieve activities related to the domain. The proposed data quality methodological approach will be based on a number of key components (later explained and justified in the thesis) namely: 1) how to identify the most important data quality dimensions of EHR Big Data in order to focus data quality initiatives, 2) how artificial intelligence and machine learning can most effectively help with detecting data quality issues in Big Data and 3) how to most effectively perform data repairs in terms of maximising computing performance and minimising reliance upon users.

1.1.1 Background

Big data has been reported as an evolving quiet revolution since a number of years. The use of data by companies has been increasing in size such that in some cases, data sources for Big Data are measured in the order of petabytes. Retail organisations like Walmart and Tesco handle millions of customer transactions per hour. Billions of people around the world work with different types of data through their mobile devices including phones and other smart devices (Majed, 2016). This increased use of data is the very premise of the term Big Data (Bollier, 2010). Moreover, with the increased use of networks, sensors, transaction processing systems and social media amongst others, organisations are facing a deluge of data which is estimated to reach a staggering worldwide volume of 40 ZB by 2020, where a ZB is equivalent to a trillion of GB (Aisling, 2013). For a specific industry like the healthcare industry, a Compound Annual

Growth Rate (CAGR) of 36 percent by 2025 is expected (Sadineni, 2020). However, there is no standard definition for the term 'Big Data'. The most accepted explanation refers to data requirements which cannot be processed by relational database and data warehouse related tools; this inability has prompted the development of a myriad of tools and techniques relative to the storage, analysis and display of data. This amalgamation of different tools and techniques gravitating around the concept of data is referred in more simple term as 'Big Data' (Demchenko et al., 2014).

The added value behind possessing such gigantic volumes, varied types and continuous production, of data resides in the capacity to analyse the data to uncover valuable ideas, make more precise predictions and understand situations. This domain of Big Data analytics is receiving a lot of research and industry attention but is also an area incurring huge inefficiencies and challenges. The use of data analytics in the field of Information Systems (IS) has been present for several years with systems such as 'business intelligence' and 'data mining'. Unfortunately, data analytics tools and techniques currently being used face stiff challenges due to the following characteristics of Big Data: *volume, velocity, variety and veracity*. The velocity aspect refers to the speed with which data produced and collected is analysed such that timely use is made out of it whereas variety refers to the different formats, ranging from structured to unstructured, of data being collected and analysed. Veracity is one of the most prominent characteristics for the purpose of this research as it is directly related to the notion of the quality level of data being used. The main argument here is that the quality of data used in Big Data systems need to conform to a certain quality level to produce an adequate quality of information, knowledge, intelligence and insight.

Making use of quality data or data 'fit for purpose' is obviously very important in order to produce actionable decisions, insights, knowledge and even intelligence from information systems following the 'Garbage In Garbage Out' (GIGO) paradigm. The data being collected or captured from different data sources might suffer from a wide range of possible issues. Those problems result in a lower quality of data and transitively have a negative impact upon the value derivable. This rationale prompted a lot of research studies aimed at improving the quality of data before being used by IS. This field or domain is often referred to as 'data pre-processing' activities and consist of data cleaning, data transformation, data integration and data reduction as main activities. Unfortunately, data pre-processing could reduce the response time and overall efficiency of the whole IS (Vattulainen, 2015). The data pre-processing tools and techniques

applicable to improve data quality in traditional data systems are currently seriously challenged by the advent of Big Data, which brings about new characteristics related to data use. The data quality community is calling out for more appropriate tools and systems aimed at addressing the veracity characteristic of Big Data (Juneja & Das, 2019; Shrivastava et al., 2019). As this involves a series of stages and layers, a framework, or more precisely, a data quality methodological approach can guide data quality initiatives for Big Data.

The health industry is a huge producer and also a big consumer of data. One example of a Big Data initiative in health industry is the "Pittsburgh health data alliance" (Pittsburgh health data alliance, 2016). This is a collaboration between Carnegie Mellon University (bringing expertise in computer science and machine learning), University of Pittsburgh (bringing expertise in medical research) and UPMC Enterprises (bringing deep data and successful commercialization experience). The aims of this collaboration include improving the level of medical solutions through data mining, lower treatment costs and produce new treatment protocols (Pittsburgh health data alliance, 2016). Another well-known Big Data application in health industry is the partnership between Apple and IBM (Marr, 2015). This partnership involves the use of IBM's Watson health cloud analytics service to power machine learning natural language computation of billions of health data items being captured by Apple's 'healthkit' development applications found on Apple watch, iPhones and iPads. The emergence of wearable devices such as 'Jawbone' (Seppala, 2015) and the increasing popularity of telemedicine and personalized medicine are pushing towards a more intelligent use of untapped health data. Health data can be categorized differently such as electronic health records, administrative data, claims data, disease registries, health surveys and clinical trials data. Google Flu and Ebola forecasts systems have been highly mediatized ways Big Data have been applied in the health industry. There is other lesser-known ways Big Data are being used in the health industry, such as:

- Big Data is used to improve decision making in the health industry by increasing the potential of EBMs "small data" (Handler, 2012). Evidence Based Medicine (EBM) is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (Sackett et al., 1996). Personalized Decision Support System (PDSS) is enhancing personalized medicine or EBM through big data analytics (Yesha et al., 2014).
- Health industry frauds are very serious issues in many countries. In UK, it is estimated that the amount of money lost to frauds amounted to £5bn in 2014 (BBC, 2014). Big Data through the help of data mining combined with machine learning can play a major role in

fraud detection (Syntelli Marketing, 2020). Big Data analytics can identify some fraud as soon as it happens and help in its prevention.

- Big data analytics are being applied with the aim of reducing patients' readmission numbers. Patient readmission is not only very expensive for hospitals, but the ratio of patients' dying after readmission is alarmingly high (Zolfaghar et al., 2013).
- Big data is proving to be a very useful tool for medical research. As there are many large medical datasets such as the human genomic dataset, pharmaceutical companies are harnessing the power of Big Data analytics to discover new medicines and understand diseases (Tett, 2013).
- With the Internet of Things (IoT) in the health industry, data is retrieved in a pervasive manner and processed in a timely manner with the data being shared across networks (Xu et al., 2014). The data collection through IoT not only retrieves data but governs the daily life of the patient. Through integrating Big Data and IoT to health services, both patients and health facilities cut down costs by reducing repetition of tests and benefit from more accurate diagnosis.

1.2 Problem definition

1.2.1 Data Quality for Big Data

Data quality applicable to Big Data in general is a relatively under-researched topic with differing schools of thought pertaining to its importance. Even a general definition for data quality is hard to find. However, the general conception is that data is of high or good quality if it is 'fit for purpose'. However, some definitions are very focused and restricted, such as limiting high quality data explanation to a dimension such as accuracy only (Hermans, 2009). Understanding more precise meanings of the concept of data quality is therefore extremely important to address research gaps in this field.

Increasing regulatory activities and an increase in the understanding of the value of data have raised the importance of data quality as a discipline across organisations. The data quality approaches which had been effective with traditional data, characterised by very stable and close ended technology such as data warehouses, will need adaptations to be effective with Big Data. This is principally due to the 'beautiful chaos' nature of Big Data, since working with such huge volume, mixed structure and ever-changing data values is sometimes perceived to be technologically daunting (Caballero et al, 2014). Volume of Big Data is expected to have a high

impact upon data quality initiatives (Taleb et al., 2016), but the variety and velocity of data causes serious challenges to the performance and reliability of data quality activities for Big Data.

Data quality dimensions (DQDs) denote a particular notion or characteristic of quality. Traditional DQDs such as timeliness, availability, accuracy, precision, consistency, security and accessibility might need to be re-considered with the specific features of volume, velocity and variety associated with Big Data (Malik, 2013). For example, data coming from sensor sources need to have the timeliness and accuracy characteristics whereas related and similar data coming from social media sources might not possess the same degree of accuracy. The Canadian Institute for Health Information (CIHI) has identified accuracy, timeliness, comparability, usability and relevance as the main DQDs (CIHI, 2015) for the healthcare sector *but these DQDs were not focused on Big Data*. On the other hand, completeness, correctness, concordance, plausibility and currency are referred to as the main DQDs for electronic health records (EHR) (Weiskopf & Chunhua, 2013), *but the results were not targeted in Big Data contexts*. One of the primary outcomes of this research study is to identify the most important DQDs for Big Data in the health industry, and to use these findings to guide further data quality activities.

Supporting data quality improvement activities relative to a set of DQDs could arguably be accomplished with the support of artificial intelligence(AI) and machine learning (ML) algorithms. High quality data depends upon the purpose and context of the use of data. Therefore, a level of flexibility and adjustments is required when data quality detection is undertaken. ML algorithms may be useful in this precise context. ML algorithms can be categorised as either supervised learning such as Support Vector Machines or as unsupervised learning such as K-means algorithm. Supervised learning models need *a priori* information about the data to build training sets, better known as a label in the ML domain (Negri et al., 2011). Statistically based generative data models are well established unsupervised learning models applied in data pre-processing tasks for data quality management activities. An example is 'BayeSwipe', which is a tool based upon Bayes Theorem to statistically predict occurrences of dirty data in Big Data (Sushovan De, 2014). However, the efficiency of this technique in detecting and correcting incorrect data is just around 40%, which is quite limited. There is a lack of clear knowledge about whether machine learning algorithms could be effective towards detecting and improving DQ in the context of Big Data for the health industry. Thus, this research study also focuses on establishing to what extent the use of AI and machine learning algorithms might support detection of dirty data as part of data driven DQ strategies.

Data cleansing activities are normally the culmination of any proper data quality management process. The following are the major challenges which Big Data seems to bring for data cleansing:

Repair upon original data: How some datasets are intended to be used is critical towards understanding the way cleansing activities would need to be carried out. Thus, in situations where the same dataset could be subject to different types of analysis, repairing or providing edits on the original dataset might prove very beneficial for a particular use case but could cause the dataset to be unsuitable for other use cases (Soares, 2012). This results in the need of data replication which might be a problem for Big Data systems. Furthermore, data repairs might result into too computationally intensive data analytics systems due to possible user interventions needed.

Computational complexity: Some data cleansing activities might involve huge processing power (Khayyat et al., 2015). For example, performing resemblance measure might perfectly be achievable in relatively smaller datasets of thousands of records with tens of attributes, but when it is being scaled to Big Data proportions of millions of records across hundreds of attributes, it could result into disastrous response times if the data quality solution is not being properly powered by due processing capabilities. Having automated techniques to deal with data quality issues in Big Data is cited to be one big challenge (Rao et al., 2015). Partly automated data cleansing techniques targeting the most important or top-k strata of data could generate the benefits of reducing the computational complexity and also being less dependent upon user defined rules.

This current research study aims, amongst others, at proposing steps to improve methods of cleansing data compared to existing tools and algorithms as identified through literature review. The new steps proposed will be experimented through a prototype, which will be benchmarked against the existing tools and algorithms.

As there are many related but distinct research gaps or emerging research areas for this research study, as addressed in the above paragraphs, it is essential to scope the research properly. Generating an approach to apply data quality in general seems highly utopic, as there are too many factors and constraints to consider. This research study focuses upon one area where the use of Big Data is promising and where the importance of data quality is high for scoping this

research, and the choice is upon the health industry. Further investigations performed during the course of the research study ascertained additional scoping of the research is required in terms of experimenting with Electronic Health Records (EHR) data only.

1.2.2 Research question

Having understood that data quality application is firstly, a largely contextual process and secondly, a still immature and evolving domain of Big Data, this research aims at answering the following main research question:

What could be an optimized methodological approach to enhance DQ in EHR Big Data ?

As would be elaborated further in the literature review, it is now clear that data quality is considered crucial for Big Data. Catering for data quality was traditionally considered very important for data mining applications and this resulted into a very tedious series of data pre-processing activities (Corrales et al., 2018). It is now believed that even for Big Data analytics, there will be the need to optimize data quality activities (Taleb et al., 2016; Corrales et al., 2018). However, the approaches, methods and techniques to perform data pre-processing activities might not be similar to those used for normal data due to the characteristics and unique challenges offered by Big Data. Therefore, developing a data quality methodological approach, consisting of precise and justified steps applicable in the context of EHR Big Data can be helpful for data quality industry players as well as research persons in the domain. This might be used as a guide to optimize steps and activities for data quality improvement in the context of EHR Big Data.

1.3 Aim

This study aims to propose a data quality methodological approach to enable the attainment of optimum data quality for EHR Big Data. A methodological approach is explained to be a series of steps to achieve a certain aim (Rahimi et al., 2016).

In that context, the first area of investigation concerns determining what are the most important data quality dimensions to focus upon. Then, AI and machine learning support for data quality activities would be investigated, for the detection of dirty data based upon some specific data quality dimensions. Ultimately, more knowledge relative to which data repair algorithms have

the best computational complexity and correct data repairs ratio relative to existing algorithms will be derived.

1.4 Objectives

A methodological approach to data quality optimisation is a precise process for implementing data quality initiatives. Based on an analysis of existing literature, three main objectives are identified (below) in developing this methodological approach. Accomplishing these objectives in the context of a specific area of application will then provide further details to inform the methodological approach. The objectives are:

(a) To determine the most important DQDs of EHR Big Data.

The whole process of optimising data quality in any kind of dataset starts with the capacity to determine 'clean' from 'dirty' data. There should be criteria or characteristics which describe data quality, which are known as DQDs. There is a lack of clear agreement from existing research concerning what are the most important DQDs in the specific context of EHR Big Data. This research aims to understand and analyse which DQDs are most important by applying well accepted research methods. This knowledge of importance of DQDs would be useful for future endeavours aiming to achieve a high level of data quality. Experts in the field of Big Data have agreed that it could be unrealistic to correct all errors (Serhani et al., 2016), hence knowing the most important DQDs would guide data quality initiatives in terms of prioritizing the characteristics which quality data must possess. Therefore, the process of understanding most important DQDs will be the first component of the proposed data quality methodological approach.

(b) To investigate appropriateness of AI and machine learning algorithms for detection of DQ issues for EHR Big Data.

Building upon the knowledge of the most important DQDs for Big Data in the health industry, the subsequent step is to be detect DQ issues according to the discovered DQDs. In the context of Big Data, this is impossible to perform manually or based solely upon human interventions, which would result in unusable system performances. Traditional linear programming-based data quality algorithms would also be inadequate due to the high computational complexity of the

task. Hence, the above two reasons indicate that the discovery of dirty data in big datasets should be based upon dynamic and incremental learning by computer systems themselves and therefore applying AI/machine learning seems to be an evident choice. However, there exists a wide array of AI/machine learning algorithms, each with their own properties, advantages and disadvantages. Thus, the second objective of this research is to investigate which AI/machine learning algorithm could be most appropriate for detecting data quality based upon DQDs discovered for EHR Big Data. The knowledge gained from this step would constitute a second part of the proposed data quality methodological approach.

(c) To investigate the appropriateness of data repair algorithms for EHR Big Data.

At this point in the research, it should be possible to identify dirty data according to precise DQDs for EHR Big Data. In any proper data quality initiative, the logical following step would be to correct the dirty data to minimize their impact for possible analytics performed upon the data. Even if a minority of stakeholders in the field of Big Data have claimed that it is not worth cleaning dirty data in big datasets due to the high computational cost and low impact upon final analysis (Soares, 2012), this research tends to take a different approach because of the following reasons. Firstly, the low impact upon final analysis would be a factor of the ratio of dirty data in a big dataset, that is. the higher the amount of dirty data, the higher the possible negative impact upon future analytics. Secondly recent research in the field has claimed that improving the quality of data is critical for successful data analytics applications for Big Data (Cai,& Zhu,2015).

Data repair algorithms are numerous and specific to the types of errors most prominent in particular datasets. There is old, well established techniques such as the application of conditional functional dependencies (CFDs) to more recent techniques and/or approaches such as the one adopted by 'BigDansing'. To the best of current author's knowledge, no current research work has discussed in depth about data repair algorithm applicable for EHR Big Data. Ridzuan et al. (2019) undertook a review of some data repair algorithms and noted the flaws and limitations of each of them in the context of Big Data. However, no actual experiments involving all those tools were carried out. Hence, the third objective of this research is to investigate the most important features that data repair algorithms should possess to be judged adequate in the context of big health datasets. The identified features would be prominent in the proposed data quality methodological approach.

Current evidence or research based on the evaluation of data repair algorithms for general big datasets such as 'BayesWipe' (Sushovan et al., 2014) or 'NADEEF' (Amr et al., 2013) tend to prove that existing data repair algorithms suffer from some issues. Over dependence upon user elaboration of data quality rules is one of the main problems denoted in current algorithms (Jesmeen et al., 2018). Thus, a completely automated data cleansing algorithm would have been the perfect solution, but due to the high criticality of data for some use cases, human responsibility would be needed to determine which data to clean. Hence, the need to develop largely automated data cleansing algorithms with minimal user interference is more realistic. A similar technique known as 'Auto discovery' had been proposed as part of a Big Data Quality framework (Taleb et al., 2015) Thus the final objective of this research is to develop a prototype data repair algorithm possessing some semi-automated features and evaluate its performance against existing data repair algorithms. The prototype would be part of experiments involving other algorithms and tools and inform the proposed data quality methodological approach.

1.5 Summary of Chapters.

This thesis is broken down in a logical and coherent way with the aim of explaining to the reader the creation of the proposed data quality methodological approach for the attainment of optimum data quality for EHR Big Data.

Chapter 1 presents a general background of the main elements related in this research in terms of Big Data, data in the health industry, the issues promoting this research in terms of the knowledge gaps existing in the fields of Big Data and data quality and the aims and objectives of the thesis.

Chapter 2 focuses on literature review and gives the current state of knowledge of a variety of different concepts forming the background of the current research study. As there are many ideas discussed, this chapter has been fragmented into two main sections; first one discussing ideas which influence the general understanding of the domain of the research, but not directly impacting it. The second section focuses more upon ideas which are considered to have a direct impact upon the potential components of the proposed data quality methodological approach.

Chapter 3 focuses on methodology and discusses different research methods which are used for this research. These methods are instrumental to derive research results, confirming hypothesis and deriving conclusions, and hence, will be prominent in the proposed data quality

methodological approach for this current thesis. The rationale towards why certain research methods and experiments have been chosen, how they are planned to be implemented and evaluated are discussed.

Chapter 4 tackles the first research objective of this research study; the determination of the most important DQDs for EHR Big Data. This chapter is architected as a journal article, in terms of its introduction, specific focused literature review, discussions about both Inner Hermeneutic Cycle and Latent Semantic Analysis, main findings and results.

Chapter 5 discusses about the second main research objective of this thesis; the determination of the most appropriate machine learning algorithms to detect data accuracy and completeness issues. This chapter is also architected as a journal article, with its own introduction, literature review focusing upon machine learning algorithms and models which had been used to solve data quality issues according to data accuracy and completeness, discussions about tools and resources used to setup the experiments, the implementation of the different experiments, discussions of successes and failures of machine learning algorithms as given by the literature review, discussions about the evaluation of the different algorithms which had been implemented relative to well established evaluation measures and criteria and the conclusions.

Chapter 6 discusses the viability, strengths and weaknesses of data repair/transformation tools. It considers both techniques and tools discussed for the improvement process of DQ lifecycle coming from existing literature, but also surveys and compares existing DQ tools which are available off-the-shelf. The focus will be upon improvement for Big Data, as there are unknowns about how well data improvement is for Big Data specifically. This chapter is also architected with its specific literature review, findings, discussions and finally, conclusion.

Chapter 7 presents and justifies a proposed data quality methodological approach which might be applicable in the context of EHR Big Data. The knowledge learnt from chapters 4, 5 and 6 are used to inform this proposal. Prior to this, a discussion around what is currently known about existing frameworks and methodological approaches is carried out. Evaluation of selected DQ approaches is made based upon validity for Big Data and the main steps as proposed through the problem definition section.

Chapter 8 summarises the key findings of this thesis. The important conclusions are highlighted and presented in a clear and coherent format. The aim is for the reader to understand what drove certain investigations, how they were carried out and the main results generated through them. It would also highlight key challenges in the given field. Limitations of the current work will be explained, and what had been attempted to tackle them and why they failed. Future areas of work will also be expanded.

The rest of the thesis contains references consulted and full source codes of different experiments carried out in the context of this thesis. This will be extremely useful for other research to build upon the current ones and would hence help the reproduction and improvement of results.

Chapter 2: Literature Review

This chapter aims to provide the current existing state of knowledge of concepts and topics which are associated with the current research. It is broadly segmented into a review of relevant background and a review or relevant related research. As such, ideas which are not central to the research such as big data technologies are probed in depth as they do inform about concepts and ideas which influence the research study. However, other ideas and concepts directly related to the research such as existing data quality dimensions, machine learning techniques and data cleansing algorithms are described and analysed within the perspective and scope of this research.

2.1 Review of relevant background

2.1.1 Big Data

One of the most important contexts of this doctoral research is that data quality is being investigated in the context of Big Data. The term 'Big Data' itself is reported to have been coined based upon a research work undertaken by Doug Laney for a Gartner institute report back in 2011. In that report, the author focused upon the main characteristics used to describe Big Data, that is, high volume, variety and velocity of data. However, it is very important for the validity of the research study to understand more precisely what Big Data is and how it might be different from 'small' or 'normal' data. Due to the inclusion of the term 'Big' in Big Data, the popular belief is that volume is the distinguishing characteristic describing Big Data. Another point of view based upon extensive review of Big Data implementations points to velocity as being the most distinguishing characteristic (Kitchin and McArdle, 2016). Kitchin and McArdle (2016) explored 26 datasets across 6 main use cases and came up with clear parameters of what constitute Big Data. The 3 classical characteristics are discussed in a more granular level; Volume is further probed and described into number of records in a dataset, storage required per record and total storage required. Hence, *some Big Datasets are not measured in terms of petabytes of total storage of data, but still they are acknowledged to be Big Data*. The fact that storage capacity is no longer growing exponentially and input/output transfer rates not matching Big Data growth might limit the volume of data organisations will choose to store (Dave & Gianey, 2016). Velocity is further decomposed into frequency of data generation and frequency of handling, recording and publishing data. Many examples of what was thought to be Big Data, such as national censuses do not qualify due to their low frequency of both generation and publishing data. Variety is decomposed into the different types of data (structured, semi-

structured and unstructured). Out of those three main characteristics, it is only velocity which is normally seen as the key distinguishing factor between small and Big Data. Velocity of data is often practically represented in terms of data streams. However, data streams are used since a long time, prior to the start of the Big Data age, which is around 2006. As the current research study is focusing on EHR Big Data, datasets of at least one million records are used as this a frequently used representation of Big Data across contemporary research (Kitchin and McArdle, 2016).

2.1.2 Big Data implementations

The following are widely cited database technologies associated with Big Data:

Google Map-Reduce is the apparatus for taking care of Big Data used by Google (Lin et al., 2015). Inputs and outputs are described in the key/value pair notation. A hidden framework parallelizes the calculation crosswise over expansive scale of commodity machines and oversees keeping up viable correspondence and the issue of execution. The Map capacity in the expert nodes takes the inputs, dividing them into little sub-issues, and takes them to operational centres (Lin et al., 2015). Each operational node can perform mapping many times, making a multi-level tree structure. In the Reduce capacity, the root node takes the outcomes from the sub-problems and unites them to get the answer of the whole problem. Widely cited issue with Google Map-Reduce refers to mainly to fault tolerance as a very large quantity of hardware could be involved and therefore the need to manage worker and master machine failures.

Hadoop [Hadoop Apache Project] is an open-source system that permits overseeing dispersed manipulation of large amount of data over groups of commodity computers utilizing straightforward programming models (Helbing & Balietti, 2011). It might scale up from individual servers to a huge number of machines, each of them offering nearby calculation and memory. Hadoop was roused from Google's MapReduce and Google File System (GFS) and eventually it has been acknowledged to be embraced in an expansive scope of occasions. Hadoop is intended to sweep big datasets to deliver results through a dispersed and very adaptable clump handling framework. Hadoop is made of the Hadoop Distribute File System (HDFS) and of MapReduce (Karloff et al., 2010). The programming model is competent to identify failures and tackle them naturally by running projects on different servers. HDFS permits applications to be kept running over numerous servers, which have for the most part an arrangement of cheap internal disk drives; the likelihood of the use of regular equipment is another favourable position

of Hadoop. A comparative and fascinating arrangement is HadoopDB, proposed by a gathering of scientists at Yale. HadoopDB was created using a cross breed association that blends the best components of two specialized arrangements: parallel databases in execution and productivity, and Map-Reduce-based framework for versatility, blunder resilience, and tractability. The essential thought behind HadoopDB is to utilize Map-Reduce as the corresponding layer over different centres running single-node DBMS cases (Abouzeid et al., 2009). Questions are communicated in SQL, and after that rendered into Map-Reduce. In particular, the arrangement actualized includes the utilization of PostgreSQL as the database layer, Hadoop as the corresponding layer, and Hive as the interpretation layer (Abouzeid et al., 2009).

Hbase is dispersed database expand on top of the HDFS as specified previously. Hbase utilizes a Log Structured Merge Tree approach: first it gathers all overhauls into an exceptional data structure on memory, and after that, occasionally, flush this memory on disk and making another file sorted out of data records. These files are unchanged after some time, while the few records added are intermittently combined. HBase's execution is attractive as a rule and may be further enhanced by utilizing Bloom channels (Borthakur et al., 2011). Both HBase and HDFS frameworks have been developed by considering versatility as a central guideline. Principle challenges faced by Hbase are internal failure handling capacity of the individual nodes.

Hive [Apache Hive] is an open-source information warehousing arrangement on top of Hadoop (Vohra, 2016). Hive has been arranged with the goal of looking at a lot of information in a more efficient way, enhancing the inquiry capacities of Hadoop. Hive supports inquiries communicated in a SQL-like revelatory dialect - HiveQL-to concentrate data from various sources, for example, HDFS or HBase. Hive is additionally described by the comportment of a framework inventory (Metastore) containing diagrams and insights, which are valuable in operations as data investigation, question enhancement and review solution. In Facebook, the Hive distribution centre contains a huge number of tables and stores more than 700TB of information and is being connected widely for both bookkeeping and specially appointed investigations by more than 200 users for each month (Thusoo et al., 2010).

MongoDB is an archive arranged database that records data in BSON, a twofold JSON group (Maktoubian,2019). Its fundamental thought comprises the use of a more adaptable model to supplant the fantastic idea of a "column". MongoDB is open-source and it is without diagram,

i.e., there is no altered or predefined record (Borthakur et al., 2011). Keeping in mind the end goal is to recover information, impromptu questions in view of these records can be utilized. Inquiries are made as BSON articles to make them more proficient and are like SQL questions. MongoDB bolsters MapReduce inquiries and nuclear operations on individual fields inside of the record. An applicable favourable position of MongoDB is store polymorphic information effortlessly, and the likelihood of making versatile cloud frameworks given its scale-out outline, which expands convenience and designer adaptability (Maktoubian, 2019). In addition, server expenses are fundamentally low because MongoDB arrangement can utilize modest equipment, and their level scale-out structural engineering, can likewise lessen capacity costs.

There are other tools connected with Big Data such as Google BigQuery, Apache Spark, Apache Storm and IBM BigInsight, amongst others. They are not being discussed in depth as this section aims to provide an overview of characteristics of some well-known Big Data technologies. These technologies can be part of the data source element related with any data related project, including data quality activities. This research study has considered the need to use any of those tools for the experiments associated with the application of machine learning algorithms to detect DQ issues and data repair experiments. The likely impact of using those tools upon the research study and the proposed data quality methodological approach has also been investigated as part of experiments. However, as the research study is an applied one which involved real life datasets or data repositories, the prevalent research method of experimenting with Big Data was adopted. The method involved experiments with very large CSV based datasets, but as part of the data quality methodological approach proposed, the capacity for data quality activities to cater for streaming data is proposed. Thus, the cloud-based data repositories and computing power were also used to represent Big Data use as part of experiments for Chapters 5 and 6.

2.1.3 Data quality dimensions in the healthcare industry

Collection of data in the healthcare industry, ranging from administrative records to the numerical values of laboratory result, is complex and challenging as healthcare data are very heterogeneous in nature (format, data collection methods, standards used). Those data are used to fill important knowledge gaps in healthcare, including improving research practices, reducing costs, increasing quality of treatments and effectiveness on medical interventions (Ginsburg et al., 2009).

The value of clinical data as a transformative agent in the U.S. health care system has resulted into six DQDs being reported most insistently across US healthcare literature. These are timeliness, equitable, care-safe, patient-centered, effective and efficient (Anderson et al., 2006; Byrd et al., 2013). Whereas timeliness, effectiveness and efficiency are considered traditional data quality dimensions, equitability, care-safe and patient-centred could be considered as desirable user quality characteristics around health data, but not really data centred quality properties.

In March 2007, the UK Audit Commission distributed a structure to bolster change in data quality. The system proposed for electronic health records (EHR) includes six key attributes (dimensions) of good quality data: Accuracy, Validity, Reliability, Timeliness, Relevance and Completeness. The structure was adjusted according to the hierarchical structure proposed by the creator of the Canadian Institute of Health, adding one more dimension in terms of integrity. As there is a very strong correspondence between data quality dimensions mentioned in UK audit commission and the Canadian Institute for Health Information (CIHI), this research study elaborates further in the next section.

The CIHI is experiencing significant growth in terms of data handling and has prompted an enhancement of their 2005 version of the CIHI Data Quality framework. This update resulted in the 2009 version of the CIHI Data Quality framework assessment tools which consists of 61 criteria, 19 quality characteristics which were finally grouped into five data quality dimension, Accuracy, Timeliness, Comparability, Usability, and Relevance. The 2009 version of the CIHI Data Framework was produced to help give a better understanding to staff handling data and to complete the data quality assessment report for their data holding regarding the uniqueness in health personnel, health expenditure, drugs, medical equipment, home and continuing care with respect to clinical data holding, which was mainly the basis for previous 2005 CIHI data quality frameworks (CIHI Data Quality Framework, 2009).

The New Zealand Ministry of Health revealed that organisations were becoming more dependent on data and virtually everything in the modem organisation does depend on data and this triggers the use of a large volume of data (Langley et al., 2006). The Ministry of Health noticed the growing demand for data quality framework to be developed for the assessment of data quality within the organization (Langley et al., 2006). The NZHIS (New Zealand Health Information Service) carried out an analysis through a survey amongst data users from across the Ministry. An open-ended question was used for the survey giving participant free will to answer questions

in the area of contextual and historical information of data collection within the ministry. The outcome of analysis reveals the inconsistency of information. In most cases, respondent had a different understanding of data collection which is contradictory. In terms of national data collection, most data are known by short names. Thus, in a scenario where National Provider Index (NPI) and Health Provider Index (HPI) appeared to some respondents as the same while to others it meant two different indexes, causing confusion. Due to the inconsistency of data management, a workshop was held in respect to NZHIS. The workshop's primary objective was to generate a data quality framework that will meet consistent and accurate assessment in all national data collection across the ministry of Health, as well as improve decision making and policy development in the health sector. The five data quality dimensions of NZHIS Accuracy, Timeliness, Comparability, Usability, and Relevance were adapted from the CIHI's data quality framework. Furthermore, the NHIS felt that there was a need for transparency and explicitness in the quality dimensions, which brought the implantation of additional data quality dimensions of 'Privacy and security' by the Ministry's Senior Advisors. These dimensions were generated basically to address the standards, legislation, policies and processes with the idea of supporting the privacy of individual information within national collection. In conclusion, the six data quality dimensions advocated by NZHIS are Accuracy, Timeliness, Comparability, Usability, Relevance and Privacy and Security.

In 2004, the Irish Department of Health and Children perceived that its existing data quality structure appeared insufficient to cater for the complexity of information systems needed in the current healthcare industry. Dermot Smyth (2004) pointed to the bitty state of the technical architecture of Irish healthcare system, and the insufficient standard and modern technique in healthcare processes in Ireland and called for system integration in the sector. In 2001, the program for change in the Irish Healthcare system acknowledged the significance of data quality questions. Gnesotto and DeVogli (2003), outlined a framework of four standard goals for the Irish Healthcare system which consist of: Fair Access, Better Health for Everyone, Responsive, and Appropriate Care delivery, and High Performance. To achieve each of the four goals, the effective use of information and a quality data dimension is fully required. The above authors concluded that the seven main data quality dimensions were Accuracy, Completeness, Legibility, Relevancy, Reliability, Timeliness and Validity.

The above discussions of application of DQDs in healthcare systems for different countries denote that the area of DQDs is quite well investigated. However, those discussions refer mostly

to cases where Big Data is not involved. Furthermore, different healthcare systems discuss different DQDs. Therefore, this work investigates whether the same standard DQDs can also be applied to Big Data and aims to determine which DQDs are most important.

2.1.4 Types of health data

It is an undeniable fact that the health industry makes use of a huge amount data with various types of complexity (Gluck, 2020). The US National library of medicine contains many publications collectively containing discussions on millions of health and biomedical concepts, synonym names and their relationships. It includes over 150 categories of codes and classifications which form the source of its 'metathesaurus' (Anon., 2013). This 'metathesaurus' has been devised to allow system developers to have a standard definition of source data which could be used by potential software and applications.

Health data are *structured, unstructured and also semi-structured* (HDK, 2014). Structured data are discrete coded values such as codes of some diagnosis, for example, 4548-4 being the LOINC code for a Haemoglobin test. They would also refer to values such as patient's names or contact numbers, which are all discrete values. Unstructured data does not have discrete and well bounded values; an easy to grasp example is the written text a doctor scribbles as part of a medical diagnosis. In a recent past, all clinically captured data were all unstructured, hence easy to understand by human beings but difficult to interpret and store by computer programs. With the increasing adoption of EHRs, a lot of previously unstructured data are being converted into either structured or semi-structured formats. Semi-structured data is a mix between structured and unstructured data. Most interfaces of health software and applications would normally allow some semi-structured mode of data capture to allow quite standardized data storage and facilitating data analysis, but also some unstructured data input so that unexpected data and knowledge could be captured, and hence increasing the value of the software.

This research study focuses more on structured data quality, as the real-world dataset used as part of the experiments contain structured data. Furthermore, most EHR related datasets are made up of structured data. Table 2.1 below details different types of structured data, together with potential standards which they adhere to, examples of the data and use cases.

Table 2.1 : List of structured data (HDK,2014)

Type of Information	Coding System/Standard Vocab	Sample Data	Used for	Remarks
Procedure	CPT (Current Procedural Terminology)		Billing outpatient and inpatient	
Laboratory	LOINC (Logical Observation Identifiers Names and Codes)	19254-2 11556-8	Laboratory, Medical Record	
Medication	RxNorm and RxCUIs		Pharmacy, Medical Record	An ontology of several vocabularies.
Diagnosis	SNOMED CT(Systematized Nomenclature of Medicine--Clinical Terms)	simple chronic anaemia (disorder) pneumonia due to Staphylococcus aureus (disorder)	Problem Lists, Medical Record	Assigned by provider
Diagnosis	ICD-9-CM	250.01 V90.83	Billing, some research and population health.	Being phased out in 2015 in US. Codes are usually assigned by professional medical coders after reviewing the health record.
Procedure	ICD-9-CM	44.31 76.0	Billing, some research and population health.	Being phased out in 2015 in US. Codes are usually assigned by professional medical coders after reviewing the health record.
Procedure	APC (Ambulatory Payment Classification)	0370	Billing	Computed based on CPT and HCPCS
Diagnosis	MDC (Major Diagnosis Category)	17 08	Billing	Computed from ICD-9-CM or ICD-10-CM
Diagnosis	Diagnosis-related MS-DRGs (MS-DRG, etc.)	69 242	Inpatient Billing	Computed from ICD-9-CM or ICD-10-CM

Procedure	HCPCS (Healthcare Common Procedure Coding System)		Billing outpatient and inpatient	Assigned by coder and sometimes clinical staff.
Medication	NDC (National Drug Code)	0067-6238	Pharmacy, Medical Record	Single code represents drug, strength, dosage, route, packaging, etc.
Diagnosis	ICD-10-CM	Z77.22 I21.02	Billing, some research and population health.	Used Worldwide except in US. Codes are assigned by medical coders.
Procedure	ICD-10-CM	7W02X0Z B2230ZZ	Billing, some research and population health.	Used Worldwide except in US. Codes are assigned by medical coders.

Analysing Table 2.1 denotes a wide array of types of data, only in structured format. The more the amount and complexity of the data, the greater the possibility of DQ issues. This informs the current research study of the complexity of carrying out generic DQ activities for an industry.

2.1.5 General use of data in the health industry

The use of data in the health industry spreads across various sectors and stakeholders including administrative staffs, clinical, users/patients, government, social care staffs and researchers. The following discusses the potential use of data by certain stakeholders.

- **Users/Patients:** In this context, patients are people whom health related services are rendered to. These users need information about their health status to be able to make informed decisions. This information is usually provided by doctors based on diagnosis and treatments carried out.
- **Clinical Staff:** Clinical staff in the health industry use data recorded in healthcare records to ensure adequate provision of services. A typical example would be patient records.
- **Social care staff:** compared to the above mentioned, social care staffs are often external to health organisations. They need information to provide general services such as community development service. For example, social care staff can gain information about children in a particular region with an aim to provide health services which will be beneficial to the community.
- **Administrative staff:** in this context, administrative staff requires and use data for administrative tasks. Such tasks include managing attendance and making right preparations for an outpatient clinic.

- **Government departments:** Government departments require data for development activities. This information is used to create health/social care policy, also for provision of related funds and engage in proper planning.
- **Researchers:** Researchers require and use health related data for the purpose of analysing and interpreting causes of diseases. Through records, researchers make attempts to find out causes and prevention or cure to these diseases.

As data quality is a context sensitive domain and is commonly explained as 'fit for purpose', knowledge of data needs of different stakeholders in the health industry might influence data quality needs and levels. Furthermore, understanding typical stakeholders in the health industry might precise the need and use of Big Data.

2.1.6 Categories of health data

Upon consultation of the US Health Resources and Services Administration website (HRSA, 2019), the vastness of the amount of data being used in the health industry could be witnessed. This US health department has categorised data to help data management activities into the following:

Health care professions data: Included are extensive data for the most current year of physicians by detailed specialty and major professional activity, age, gender, and graduation location; and the most current data available for other major health professions. Aggregate physician data are available from 1970 to the present, with more detailed speciality data available for some five-year intervals. Also included are data for dentists, physician assistants, nurse practitioners, nurse midwives, nurse anaesthetists, chiropractors, optometrists, and podiatrists, among others. Additionally, AHRF contains information regarding Health Professions Shortage Areas (in Codes and Classifications).

Health professions training data: Data are provided on the number of schools, enrolments, and graduates for major health professions, including MD Schools, DO Schools, Dental Schools, Dental Auxiliary Schools, Veterinarian Schools, Pharmacy Schools, Optometry Schools, and Podiatry Schools.

Health facilities data: Included are current and historic information on the characteristics of and services offered by hospitals. Statistics include number of admissions, inpatient days, outpatient visits, beds by type, number of personnel by category, etc. Data are provided on nursing homes, home health agencies, hospices, ambulatory surgery centres, National Health Service Corps Sites and more.

Population Characteristics and Economic Data: Included are age, race, and sex for the census year populations as well as total population for intervening years. Also included are data on mortality, infant mortality, birth statistics, education, Medicaid eligible persons and Medicare enrolees including those in Medicare Advantage, in Fee for Service, and in the Prescription Drug Program. Economic data include civilian employment and unemployment; total, per capita, and median income; poverty; housing statistics; and distribution of families and individuals by income groups. Health insurance statistics and SNAP recipients are also included. New on the 2013-2014 and later releases of the AHRF are statistics regarding disabled and veteran populations.

Environment data: Included are population and housing density, land area, and air quality and ground contamination data.

Codes and Classifications: Included are geographic descriptors such as Core Based Statistical Areas; Rural/Urban Continuum Codes, Urban Influence codes, county typology codes, Federal region codes, Census county group codes, Census contiguous county codes, Health Professions Shortage Areas, among others. New on the 2013-2014 and later releases of the AHRF are Indicators of Persistent and Deep Poverty.

This section illustrates the breadth of types of data used in the health industry. This breadth increases challenges to impose high levels of data quality and increases the complexity of developing activities as part of a unique data quality methodological approach.

2.1.7 Classification algorithms

This research study aims to investigate the possible use of AI/ML classification algorithms for the purpose of detecting dirty data from clean ones. Most classification algorithms are based upon supervised learning ML algorithms. This section provides a background description of how supervised learning works, well-known supervised learning algorithms and evaluation procedures.

A ML classifier algorithm, also known as classifier, refers to a predictive modelling problem where a class label is predicted for a given example of input data (Brownlee, 2020). The examples of input data are called observations or patterns. The aim of ML classifiers is to correspond unseen examples most adequately to a potential class label. Classification tasks which are most known are binary classification, multi-class classification, multi-label classification and imbalanced classification. For the current research study, classification of dirty data is expected

to be either a binary classification problem since the two most important DQDs determined in Chapter 4 are *accuracy* and *completeness*. The dirty data will correspond to some measurement of a dimension or not. For example, we could say that a certain attribute value is complete or incomplete. Different classification algorithms are applied in the health industry in the determination of effective surgery procedure, medical tests, medication, diagnostics, and clinical data findings.

Supervised learning makes use of given datasets known as the "training set" to do predictions. Two sets of value are in the training set: the input data and the labels. The training set is used by the supervised learning algorithms to attempt at building models. A model often utilizes test datasets for validation (Brownlee, 2016). Test data are used to ascertain the utility existing in a predictive relationship and the strength of such relationship. Although the test set does not depend on the training set, they share a probabilistic distribution. When a model fits the training set but does not fit the test set, it is known as a case of 'overfitting'. If the model fits the training set and fits the test set as much, then, it is a minimal 'overfitting' occurrence (Brownlee, 2016). The use of very large training datasets might result in models that have very high predictive values. The closer the prediction accuracy to 100%, the better the model is considered, but no actual threshold is used to ascertain whether a model is applicable and the actual comparison of models is subjective depending upon the data being analysed. When a classifier that best satisfies a particular problem is needed, after the candidate algorithms are trained by the training sets, the choice of the best candidate is made using the validation set comparisons of the candidate set performances.

A validation set is used to ensure that there is no 'overfitting' as classification observations are being adjusted. The validation dataset is used to estimate prediction error for model selection; in the case of data classification algorithms, it is this dataset which is going to be used to rate the efficiency of the training of specific machine learning algorithms from the training sets. In the Holdout method a proportion of the training dataset is kept apart to be used as the validation dataset. The proportion of training to validation sets is usually 70% to 30% (Gareth, 2013). An alternative process is to do a cross-validation which is to repeatedly partition the initial training set into training and validation sets respectively. The repetition of the partitioning could be done in many ways, example is to divide the training set into two halves, one half is used as the training set while the other is used as validation set and then the role is alternated for the halves afterwards. A randomly selected subset could also be used as the validation set.

The term supervised learning simply points to the fact that some given sets of data are being analysed to predict a discrete categorization of the new observations according to the given or known categories or classes of instances or observations. The alternative is prediction by using regression i.e., a prediction of continuous-response values. The supervised learning therefore comprises the common classification and the common regression algorithms (Brownlee, 2016).

There is limited number of clearly observed inductive logical studies on comparative supervised learning techniques for data quality. The most acclaimed research is STATLOG (King et al., 1995). The algorithms compared in STATLOG came from different domains: statistics such as Naïve-Bayes, symbolic learning such as CART and neural networks such as back-propagation. Amongst the twelve datasets used in that research, three were from the health industry. It was found that the performance of machine learning algorithms depended upon descriptors associated with datasets. This calls for empirically focused research involving newer machine learning techniques focused on specific types of health data and what is the current impact of descriptors for data quality tasks.

The performance criteria like the classification accuracy, F-measure or sensitivity score amongst others are obtained when the ML algorithms are applied on the test set (Mishra, 2018). Supervised learning techniques are applied in different fields with various performance and observation measurements are used as appropriate for each field. For information retrieval precision/recall is most preferable, ROC area is used for medicine while Lift is adhered to in marketing (Mishra, 2018). The performance of each technique that is excellent for one field gives a differing performance when used for some other area. As a result, it becomes necessary to use the most adequate performance metrics per different contexts or fields to evaluate algorithms.

As thoroughly described above, there are myriads of research involving machine learning algorithms, including those focusing upon measurement and metrics, already carried out. Despite all those, it cannot be clearly ascertained how far and exactly how ML algorithms could support DQ activities in the context of Big Data for the health industry. The use of classifiers is theoretically very appealing in order to distinguish clean from dirty data, and hence can be useful for DQ issues detection. It might also provide better efficiency for data repairs. However, previous studies involving ML algorithms in DQ activities is quite rare. This calls for focused

research of the possible use of AI/ML on EHR Big Data with the goal of detecting 'clean' from 'dirty' data.

2.1.8 Measurements and metrics

As briefly outlined in the previous section, the overall assessment of data quality level in a given context is closely linked with measuring the dimensions of data quality. Most metrics used for measurement of data quality are normally within a range from 0 to 1, with 0 representing incorrect value and 1 representing a correct value (Blake & Mangiameli, 2011). Many dimensions such as accuracy, completeness and consistency amongst others are calculated by the following function:

$$D = 1 - (Ni/Nt) \qquad (1)$$

Where D is the metric for a given dimension, Ni is the number of incorrect values and Nt is the total amount of values for the dimension concerned. This measurement and associated metric would definitely still hold even for Big Data, but it could be quite difficult to derive both Ni and Nt in situations where there is constant input stream of data. Thus, the velocity aspect of Big Data could be the most problematic property in terms of Data Quality measurement; but if this velocity aspect has been mastered, that is *Nt* is well established, then there is no reason why the same metrics will not be applicable for Big Data.

Since this research aims to evaluate the most appropriate DQ detection and data repairs techniques, there is the need for measurements during the experiments; taking some specific DQDs which would have been asserted to be most important for Big Data in the healthcare industry, the research would need to establish the baseline metrics. For example, a dimension like completeness needs to be measured in the original datasets we have identified for the different experiments. Metrics as described earlier would be applied as part of potential data auditing activities. After either some data detection or data repairs algorithms have been applied on the benchmark dataset, the same metrics could again be used to measure a given DQD. Ultimately, some comparison could be achieved based on the measurements done pre-experimentation compared to those done post-experimentation**,** but this will depend upon the knowledge upon the exact value of *Ni.*

2.2 Review of related previous research

2.2.1 Data quality and Big Data

Traditional ways of handling dirty data might not be adequate for Big Data due to the amount of time and resources which might be required to clean the data (Ridzuan et al., 2019). Volume but more importantly variety of Big Data are cited as the main data quality challenges for Big Data (Ridzuan et al., 2019). Manual methods of applying data quality activities and the need to use domain experts are cited to be very important elements for Big Data quality activities. Some years ago, a school of thought advocated that data quality is not really a high concern for proper Big Data analytics. The main argument put forward was that the volume and exhaustibility of data is so huge that the amount of dirty data might have only minimal effects on the results of further analytics (Soares, 2012). However, a few authors have rejected this argument and advocated the idea that improper data quality has a high negative impact upon potential future analytics and other Big Data uses (Caballero, Serrano, & Piattinni, 2014;Shi, et al., 2015; Kichin & Lauriault, 2015).

The causes of data quality issues in the Big Data context are numerous and various as is the variety of types of data that may be associated with big data. One prominent example of improper data quality concerns issues with sensor-based data. Harsh weather conditions and improper maintenance of sensors are reported to inevitably lead to sensor-based data becoming dirty, especially in the context of sensors used for power grids (Shi et al., 2015).

Data pre-processing consists of data preparation and data reduction techniques. Data preparation involves data transformation, integration, cleaning and normalization whilst data reduction aims at reducing complexity of data by feature selection and instance selection (Garcia et al., 2016). Data preparation embeds data quality improvement activities. Garcia et al. (2016) also claimed that more and more researchers in the field of data mining are adopting data pre-processing techniques as part of existing frameworks or alternatively creating entire new ones. Garcia et al. (2016) also investigated some data cleansing methods operating upon Hadoop Map-Reduce using deep analysis of missing information to deal with incomplete data. Hence, there is the start of some data pre-processing initiatives on Big Data, but without complete information on the type of data and datasets involved. Furthermore, as the work of Garcia et al. (2016) involved solving

missing data issues, the techniques involved might be highly relevant for informing the data quality methodological approach to be developed by this current research study.

2.2.2 Data quality dimensions

The notion of quality is very often expressed in terms of *dimensions*, which are different subjective ways to describe data quality for specific purposes and contexts. Throughout the literature, there have been lots of different sets of dimensions being considered by several authors. A brief comparison of research studies detailing data quality dimensions is given hereunder:

1. Pipino, Wang and Yang (2002) have been some of the most widely cited authors who have investigated how to measure or assess level of quality of data. They argued that some assessments of data quality could be task independent, *therefore not restrained by the context of application while others are task dependent*. Table 2.2 below depicts the main dimensions they thought were worthy of discussion:

Table 2.2: List of data quality dimensions (Pipino, Wang and Yang, 2002)

Dimensions	Definitions
Accessibility	Extent to which data is available, or easily and quickly retrievable
Appropriate amount of data	Extent to which volume of data is appropriate for the task at hand
Believability	Extent to which data is regarded as true and credible
Completeness	Extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Consistent representation	Extent to which data is presented in the same format
Ease of manipulation	Extent to which data is easy to manipulate and apply to different tasks
Free-of-error	Extent to which data is correct and reliable
Interpretability	Extent to which data is in appropriate languages, symbols and units, and the definitions are clear
Objectivity	Extent to which data is unbiased, unprejudiced and impartial
Relevancy	Extent to which data is applicable and helpful for the task at hand
Reputation	Extent to which data is highly regarded in terms of its source and content
Security	Extent to which access to data is restricted appropriately to maintain its security
Timeliness	Extent to which data is sufficiently up-to-date
Understandability	Extent to which data is easily comprehended
Value-added	Extent to which data is beneficial and provides advantages from its uses

***Implications for the current research**:* The current research study will endeavour to verify which DQDs are most important for Big Data in the health industry. Thus, an extensive review of past work and the identification of patterns in terms of dimensions discussed would ascertain how important the above mentioned DQDs could be in Big Data for the health industry in general.

2. Batini and Scannapieca (2006) discussed the idea that each DQD is needed to cover specific aspects which might fall under the general idea of data quality. With each dimension, *there should be several metrics which could be applied to quantify a given dimension, and for each metric, there could be more than one measurement method*. Table 2.3 below summarizes the dimensions discussed by the above-named authors:

Table 2.3: List of dimensions as per Batini and Scannapecia (2006)

Dimension	**Definitions**
Accuracy	Closeness of representation of a real-life phenomenon that a data value tries to represent; measured by edit distance comparison functions for syntactic accuracy
Correctness	Also termed as semantic accuracy, which refers to the closeness of a data value with respect to a domain
Completeness	Measure of missing values for a specific column in a table; often illustrated via the NULL value and which could represent facts as value not existing, value existing but unknown and not knowing if value exists. Influenced by the closed and open world assumptions.
Currency	Concerns how promptly data are updated; can be measured by the *lastupdated* metadata
Volatility	Characterizes frequency with which data varies in time; metric given by length of time data remains valid
Timeliness	Expresses how current the data is for the task at hand; involves currency measurement and check whether data is available before planned usage time
Consistency	Concerns the violation of semantic rules defined over data items and usually expressed as integrity constraints
Accessibility	Ability for a user to access data from his own culture, physical status and technologies available.
Believability	Whether a certain source provides data which can be considered as true, real and credible
Reputation	Considers how trustable is an information source
Objectivity	Takes into account impartiality of sources
Value added	How beneficial data is and advantages derived from their use
Relevancy	How applicable is the data for the current task
Ease of understanding	How much data is clear, without ambiguity and easily comprehended

A mere comparison of the dimensions mentioned by the two sets of authors cited above clearly indicates a high level of correlation and similarity.

Implications for the current research: The results of the study from Batini and Scannapecia (2006) raise two important issues namely: whether the above discussed DQDs are still relevant in the context of Big Data in the health industry and which of these DQDs are most important for data quality in the context of Big Data in the health industry. These issues will be further investigated in the present research study and the answer to these issues forms an angular component in the proposed data quality methodological approach.

2.2.3 Dimensions of data quality for Big Data

Big Data characteristics bring new challenges for data quality processes; the high volume and velocity properties of Big Data entails that data quality activities face considerable computing resources challenges, and at the same time, data quality activities should not hinder the value possible from Big Data analytics by increasing its cost. Also, due to data coming from multiple sources, there is a need for a higher method of data integration to harmonize the semantics of the data being used (Saha & Srivastava, 2014). Thus, implementing data quality activities are very important for Big Data, even if there is scarcity of knowledge about how best to implement these data quality activities. On the other hand, the importance of improving data quality for Big Data might not be so important as the amount of incorrect data is deemed to be negligible to affect the final outcome after data has been analysed (Soares, 2012). Thus, which of those two contrasting schools of thought is relevant seems to depend on the amount and impact of the erroneous or 'dirty' data as part of a big dataset. This increases the importance of understanding which DQDs are more important for Big Data in general case or specific context. Ultimately, as data quality might be a very cost intensive process, there is a need to investigate how to measure the most relevant DQDs for Big Data and potential trade-offs between benefits and costs. This investigation will form part of the present research study in terms of factors to consider when detecting and repairing DQ issues as part of the proposed data quality methodological approach.

Caballero et al (2014) posit that the main DQD to consider for Big Data is consistency, which they explain as the capability of information systems to ensure uniformity of datasets when data are transferred across networks and systems (Caballero et al., 2014). Their main hypothesis is that the business value of a dataset can be estimated only in its context of use. They further

subdivided consistency into three subsequent parts, as discussed here under. However, they connected many of the traditional DQDs with the three consistency subdomains as follows:

- Contextual consistency refers to how far big datasets are used within same domain of interest independently of data format, size and velocity of production of data. Thus, relevancy, credibility, ease of understanding, accuracy and confidentiality are considered to be very important for contextual consistency to occur.
- Temporal consistency conveys the idea that data needs to be understood in a consistent time slot, such that the same data might not be comparable if they are not from the same time slot. Time concurrency, availability and currency are deemed to be essential DQDs for temporal consistency.
- Operational consistency brings in the operational influence of technology on the production and use of data. Availability, portability, precision, completeness and traceability are considered the main connected DQDs for this subdomain.

Caballero et al (2014) mapped how the 3v's of Big Data affect the 3Cs of data quality as shown in Table 2.4 below:

Table 2.4: Matrix of 3Cs relative to the 3Vs (Caballero et al, 2014)

	Velocity	**Volume**	**Variety**
Contextual Consistency	Consistency, Credibility, Confidentialiy	Completeness, Credibility	Accuracy, Consistency, understandability
Temporal Consistency	Consistency, Credibility, Currentness, Availability	Availability	Consistency, Currentness, Compliance
Operational Consistency	Completenes, Accessibility, Efficiency, Traceability, Availability, Recoverability	CompletenessAccessibility, Efficiency, Availability, Recoverability	Accuracy, Compliance, Accessibility, Efficiency, Traceability, Availability, Recoverability, Precision

However, the methodology used to map the DQDs to the 3v's were based *solely upon hypotheses and no actual research method was applied* to generate this mapping. Furthermore, there is a lack of precision regarding the types of data for which the above result applies. Thus, there is a research gap in the area of Big Data quality to further corroborate the importance of consistency as the most important DQD for Big Data. This must be accomplished by applying a well-accepted and suitable research method. However, as research covering Big Data quality in general is practically unfeasible due to the sheer amount of data involved and differing industry needs, this present work will focus on importance of DQDs relative to Big Data in the health industry.

According to the UNECE Big Data Quality task team, there is a hierarchical structure consisting of three 'hyperdimensions' namely *source, metadata and data*; some DQDs are nested within each of the above 'hyperdimensions' (UNECE, 2014). The DQDs discussed are as follows:

1. **Institutional/Business environment:** this refers mostly to the *effectiveness and credibility* of the agency producing the data.
2. **Privacy and security:** the task team has included those two factors as DQDs, and due to the nature of Big Data, they recommend that those two factors are given greater prominence in Big Data quality frameworks.
3. **Complexity:** refers to the lack of simplicity and uniformity of data; thus, could be considered to be equivalent to the traditional DQDs of *'ease of use' and consistency*. Definitely, due to variety of data formats and various data sources with Big Data, this dimension could be very important.
4. **Completeness:** refers to the extent to which metadata are available for proper understanding and use of data. This is quite different from the traditional DQD discussed earlier in the present research study.
5. **Usability:** refers to the extent of being able to work with data without the employment of specialized resources. As Big Data entails using multiple heterogeneous sources of data, this dimension seems logically relevant.
6. **Time factors:** more precisely referring to *timeliness and periodicity*; due to the velocity aspect of Big Data, this dimension might be important.
7. **Accuracy:** refers to the degree which data describes real life values. However, for Big Data, the notion of *selectivity* is considered very important, which basically hints at the representativeness of the dataset. Thus, for some use cases, a dataset might have a low level of selectivity whilst the same dataset might have a high selectivity level of another use case.
8. **Coherence:** quite closely linked with the traditional DQD of consistency; here, the sub-dimension *linkability,* is of relevance for Big Data as it focuses upon the ease which data can be linked between different datasets. The second sub-dimension *consistency* here refers to the extent with which a dataset complies with standard definitions.
9. **Validity:** relates to the traditional DQD of coherence.
10. **Accessibility and clarity**
11. **Relevance**

The DQDs from the UNECE report and from the other sources mentioned previously show that there is a lack of coherence in the use of terms and jargons in conjunction with data quality. For example, the dimension 'completeness' might refer to two completely different ideas, and thus are highly subjective based upon their authors' vantage point. The UNECE report focuses on the application of Big Data by different national statistical offices and *therefore might represent a Big Data quality framework for a very specific use case*. Finally, like this current research's opinion upon the study of Caballero et al. (2014), the DQDs discussed in the UNECE report is the highly subjective opinion of the members of the task force, and *no scientifically valid methodology has been applied to link those dimensions with Big Data*.

The veracity characteristic of Big Data is argued to be of crucial importance for the health industry (Raghupathi & Raghupathi, 2014). The two main reasons forwarded are (1) inaccurate data could ultimately result in life-or-death decisions for patients underscoring the fact that accuracy is a very important DQD and (2) the high level of incorrectness usually present in doctors' prescriptions lead to a lot of wastages and inefficiencies, at the very least. Hence, correctness could be derived to be another important DQD according to the authors. Lastly, as the veracity characteristic refers to the confidence in the use of data, believability and trust could be assumed to be also extremely prominent DQDs. However, accuracy, correctness, believability and trust discussed here are *the interpretations of the current author and is not based upon any solid research method*.

There are other research studies associated with DQDs investigations in specific healthcare contexts in terms of electronic health records (EHR), but there is *no specification whether those EHRs deal with Big Data* (Weiskopf & Chunhua, 2013). The DQDs discussed and identified by Wieskopf & Chunhua (2013) are completeness, correctness, concordance, plausibility, and currency.

Implications for the current research*:* the current research study will endeavour to apply a systematic and scientific methodology to investigate the most important DQDs in EHR Big Data.

2.2.4 Data quality dimension for electronic health records (EHR)

Some authors attempted to determine which DQDs could be more adequate for EHRs and what could be potential data quality assessments for each of the targeted dimensions (Weiskopf & Chunhua, 2013). The authors applied a review of literature based upon keywords to be able to identify research relative to data quality for EHR. They made use of a wide-ranging number of terms including some well-known DQDs such as 'data completeness', 'data consistency' and 'data error' amongst others. The following five DQDs were found to be more widely cited amongst an initial list of 230 articles: **Completeness, Correctness, Concordance, 'Plausability' and Currency**. Weiskopf & Chunhua (2013) further explained that they could link other DQDs to the 5 above as their meaning seems to be overlapping (see Table 2.5 below):

Table 2.5: Links of different dimensions (Weiskopf and Chunhua, 2013)

Completeness	Correctness	Concordance	Plausibility	Currency
Accessibility	Accuracy	Agreement	Accuracy	Recency
Accuracy	Corrections made	Consistency	Believability	Timeliness
Availability	Errors	Reliability	Trustworthiness	
'Missingness'	Misleading	Variation	Validity	
Omission	Positive Predictive Value			
Presence	Quality			
Quality	Validity			
Rate of recording				
Sensitivity				
Validity				

***Implications for the current research**:* The work carried out by Weiskopf and Chunhua (2013) is highly focused on a specific application context in health industry, that is, EHRs. An extensive literature review approach was used but there was not enough information pertaining to the type of datasets from which the EHRs are extracting the data. Hence, it is *quite unclear whether those dimensions for EHRs still hold in the Big Data context*. As the vast amount of EHRs data is structured, the results of the current research study could confirm the above five dimensions or propose some other ones too. The research method adopted by Weiskopf and Chunhua (2013) can also be replicated, improved and experimented in the context of the current research study.

2.2.5 Rise of open data

Use of open data is a major source of value addition for a variety of different stakeholders due to fact that it can be very easily accessed and shared by anyone for any purpose. As all other datasets, the quality of data would affect the quality of information and the intended use out of the information. However, one of the main data quality issues for open data resides in its negative impact upon reuse of data (Vetro et al, 2016). The main DQDs which could be extracted from the above-mentioned research are as follows:

- Accuracy issues such as problems caused by bad manual transposition of zip codes.
- Aggregation or integration issues such that it was impossible to reconcile financial data of companies following merging of different organisations.
- Completeness is discussed as part of missing values causing interpretation issues.
- Timeliness in terms of currency and expiration of data values

Implications for the current research: This current research study considered some open datasets made available by different entities such as from the UK government, as part of sources of data provenance of Big Data. However, as no major DQ issues were found in the open datasets available, they were not further used for experiments. The results of the current research study in chapter 4 confirm that two of the four above mentioned DQDs are amongst the most important in the context of EHR Big Data.

2.2.6 Big Data and real time data

One important consideration to cater for when discussing Big Data is its velocity aspect. Some Big Data implementations use a data lake approach, where huge amount of data is stored in temporary storage or warehouse and analytics performed on the data lake. Hadoop seems to consider this kind of approach. Other implementations work with real time or near real time analytics of data, traditionally with the use of SQL queries that operate over time and buffer windows (Wahner, 2014). 'Live data marts' are ways to process streaming data in-memory, that is, the working memory of those data marts have a huge capacity sufficient enough to avoid sending data to secondary storage for temporary storage during processing. Typical use cases of real time streaming data are fraud detection, algorithmic trading and network monitoring. These use cases would very often also encompass the volume and variety characteristics of Big Data. In the health industry, some use cases such as insurance claims fraud could be applied in a real time streaming context in the case of online automated apps, but it could also be in terms of non-

real time or batch mode analytics if the insurance providers do not aim to provide instantaneous acceptance of claims to their customers.

Implications for the current research: As part of the proposed data quality methodological approach developed by the current research study, the use of real-time or batch mode of data can have a significant impact upon data quality activities. With real-time streaming data analytics, DQ issues detection and repairs will also need to be real-time, and hence cannot tolerate any latency involved with human intervention. This could be risky with the use of critical and sensitive data. Whereas, with a batch mode source of data, DQ issues detection and data repairs will not face this extreme time limit challenge and can therefore accept semi-automated mechanisms.

2.2.7 Information and data quality

Many research studies denote quality of data and quality of information as being synonymous (Todoran et al., 2015). This leads to the fact that data quality and information quality (IQ) dimensions were being considered identical. However, the distinction between data and information needs to be investigated in different perspectives to produce higher value addition for the users of a particular IS (Todoran et al., 2015). Todoran et al., (2015) argue that as an IS consists of several modules, the input and output of each of those modules has an impact on the final information quality. A key distinction here is that data quality refers to input quality whereas information quality refers to output quality (Todoran et al., 2015). Hence, they argue that 'accuracy' is more of a data quality dimension only, whereas 'reliability' and 'timeliness' are both data and information quality dimensions.

In their research study, Todoran et al. (2015) used a target recognition system as a proof of concept. This system is made up of several modules, such as 'radar signature classification' and 'identify friend/foe'. Each of those modules are described as having their own set of quality measures for the system to be more valuable. Hence the distinction between data and information quality is exemplified by the application in those two modules as in Table 2.6 below:

Table 2.6: Examples of distinction between data and information quality

Module	DQ Dimensions	IQ dimensions
radar signature classification	Amount, accuracy, currency	Reliability, currency
identify friend/foe	accuracy, currency	Reliability, currency

Implications for the current research: The current research study focuses solely on the 'input' aspect of IS, and hence only DQDs will be investigated as part of the proposed data quality methodological approach by the current research study. This is in line with data-driven strategies of DQ approaches which will be detailed in the subsequent sections of this thesis. The rationale is that the quality at the storage level of data needs to be appropriate, irrespective upon subsequent processing of the data.

2.2.8 Dirty data and Data cleaning methods

The data cleaning methods applicable to any type of dataset depends on the nature and type of DQ issues. There are many different possible occurrences of dirty data in typical database applications. There are several existing research studies expanding on the different possibilities of dirty data and they have been divided into the following taxonomies:

1) **Muller and Freytag's data anomalies** (Müller & Freytag, 2003). The main logic here is that data which does not conform to a certain domain constraint would be equivalent to dirty data. Examples of errors highlighted by those authors are lexical errors, duplicate records, integrity constraints violation error, missing values, missing tuples among some others.
2) **Rahm and Do's classification of data quality errors** (Rahm & Do, 2000). In this taxonomy, the authors classified dirty data into multiple hierarchies starting from single source and multi-source errors, and within each of them into schema level and instance level problems. They assert that multi-source errors are more complex and are closely related to the consistency DQD, such that the same data attribute/feature could be named and structured differently over different datasets. Furthermore, overlapping data might result into duplicate records and values. They classified 19 different types of dirty data with the most notorious ones being missing values, domain integrity related issues, word transpositions, naming and structural conflicts.
3) **Kim's taxonomy of dirty data** (Kim, 2002). This research posits three main categories of dirty data: missing data, not missing but wrong data, not missing not wrong but unusable data. Upon decomposition and analysis, the authors came up with 33 different types of dirty data including missing values, uniqueness violations, inconsistent data, outdated temporal and spatial data, extraneous data, entry into wrong fields and ambiguous data.

4) **Oliveira et al's taxonomy of data quality problems** (Oliveira et al., 2005). The categorization of errors ranges from single value issues up to multi source problems. This taxonomy came up with 35 different types of errors, but which have already been broadly covered with the previous descriptions.

***Implications for the current research**:* In this current research study, the DQ issues are clustered and grouped according to the DQDs which are more important. The new possible data repair algorithm will have to target specific types of errors grouped per DQD, as discussed in the above taxonomies. The above taxonomies of dirty data provide an overview of concrete DQ issues which might exist within data, and which might be associated with some DQDs.

2.2.9 Data cleansing

Data cleansing is a well cited process and potentially involves the highest amount of data repairs with respect to data quality activities. The need to transform or edit some data source to meet certain data quality standard is an important dilemma when it comes to Big Data quality; as the same data could be used or analysed towards different use cases with Big Data, transforming the original dataset according to the business rules for one use case might negatively impact the Big Data activities for another use case with the same original dataset (Loshin , 2014).
Some researchers from IBM-research India identified four stages as part of data cleansing process for large enterprise datasets, as summarized in Figure 2.1 below (Hima et al., 2011).

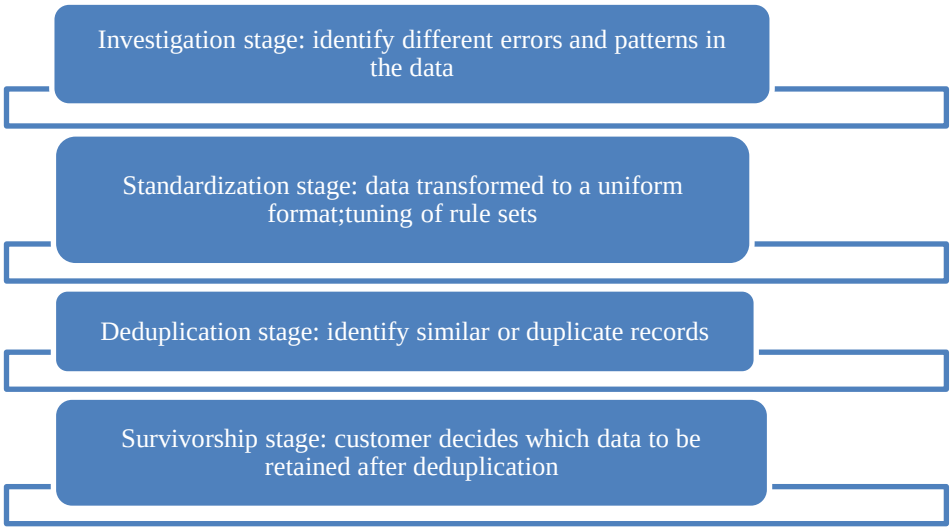

Figure 2.1: Main stages of data cleansing process (Hima et al, 2011)

The above stages are corroborated by other research studies in the field of data cleansing and/or DQ frameworks (Taleb et al., 2015; Corrales et al., 2018; Juneja & Das, 2019). To accomplish

the above four main stages, Hima et al. (2011) cite different tools and methods which should be applied as part of or supporting the different activities. The following gives a brief summary of some of them:

- **Classification:** building classes from the data in a dataset is explained to be the first step to build rule sets. An example of classification for address standardization is that the 'street' value could be linked with corresponding values such as 'Road', 'avenue' or 'lane'; thus, all these values could be assigned the same classification label. The question is whether this technique might still be valid in the context of Big Data, where the sources of data might be dynamic and varying, and therefore, potentially difficult to form finite classification labels.
- **Patterns:** they give a generalized view of how the data is formatted; it involves parsing the data and classifying all tokens into appropriate classes and replacing those classes with pre-defined labels. An example of a resultant pattern could be *N/N+B+II* where N represents numbers, B is a label for city names and I represent a single alphabet. Rules are then written to process the patterns. The use of patterns can be very useful for DQ issues detection and data repairs, and further investigation will explore how efficient and/or realistic this technique might be for Big Data in the health industry.
- **Dictionaries:** standardized data could be validated against some domain to ensure proper identification. For example, a dictionary of available cities in a country could be used to validate city names stored as part of addresses in a dataset. However, is the use of dictionaries in the context of Big Data realistic? The stages of the data quality methodological approach will explore this query.
- **Discovering variants of a term:** this is a very elaborated sub-stage which would involve the use of *reference sets* for each token or value of data; use of *syntactic clustering* which are about records which, possess the same set of terms in a sequence except some minor differences; use of *resemblance measure* for detecting groups of similar records which would use a formula to denote closeness of different records; use of a *diff-utility* to find the difference between groups of similar records. Those techniques might be very useful for the data improvement stage of the data quality methodological approach.

Thus, the Hima et al. (2011) have proposed a data-driven tool relying upon detecting characteristics of 'dirty' data in a given dataset. The dependence of using domain experts has been minimized. However, the question as to whether this same proposed data cleansing methodology could be applied to Big Data datasets is more than ever relevant.

Implications for the current research: The current research work will therefore attempt to investigate (1) Can DQ issues detection methods relying upon reference sets be realistic for DQ activities for Big Data and (2) the performance of new semi-automated data repair method which could be making use of some of the stages, methods and tools discussed above.

There are several categories of data cleansing methods. This current research study focusses further on the following ones:

Statistical methods

These methods are based on the analysis of the distribution of data in datasets or data sources. They have been used in systems proposing data quality solutions for traditional, non-Big Data systems. Examples are pattern based and quantitative error detection and repairs solutions. However, they are more apt for structured, numerical and non-categorical data, and therefore, might not be so suitable for all data types possible with Big Data. Statistical based methods are very important as data auditing techniques, even for Big Data systems as later demonstrated by the current research study.

Rule based data cleaning methods

These methods are extremely popular for data cleaning in non-Big Data systems. They rely principally upon the generalization of functional dependencies which are translated into denial constraints (Mahdavi et al., 2019). Apart from the NADEEF system, discussed further below, the KATARA system can also be considered as rule-based, and is implemented differently, even for Big Data. This efficiency is presumably due to the use of a knowledge based and crowd sourcing approach.

Machine learning based methods

Use of machine learning for both error detection and error repairs is becoming more prominent in the DQ domain nowadays. There are many tools and systems which apply different machine learning algorithms and models. Examples are HOLOCLEAN, ACTIVECLEAN, SCARE AND ERACER. Most of those systems and/or tools are further discussed in chapters 5 and 6 of this current research study.

To implement the different categories of data cleansing methods, there has been a plethora of tools which have been devised over time. A review of some of the most pertinent data cleansing methods related to the current research is discussed as follows:

BigDansing

One of the challenges with implementing data cleansing when scaling to Big Data is dealing with user defined functions (UDFs) (Khayyat et al., 2015). BigDansing is a new architecture to incorporate the application of UDFs more efficiently. The typical data cleansing steps as described by the above-named authors are (1) specifying quality rules, (2) detecting errors w.r.t data quality rules and (3) repairing detected errors. However, detecting and repairing data quality issues face some difficulties namely:

a) High complexity of rules leads to intractable computations over large datasets, thus limiting the applicability of data cleansing systems for Big Data.
b) Effective parallelization is hard to achieve with UDFs when the latter is specified using procedural languages.

'BigDansing' is reported to deal with the two difficulties above by (1) abstracting and simplifying the process of rules specification for UDFs and (2) to enable the application of distributed repair algorithms. 'BigDansing' was benchmarked to other systems which could support some level of data cleansing routines such as Spark SQL, Shark, NADEEF and PostgreSQL. The results show that 'BigDansing' outperforms the other systems using measures such as time to scale data quality activities upon large datasets, higher efficiency in deduplication of large datasets and improvements of repair efficiency.

However, it could be argued that the process of rule specification being the responsibility of users could be one of the limiting factors of a system such as 'BigDansing'.

Implications for the current research: In this current research, the application of machine learning techniques would be investigated to identify the dirty data as part of a big dataset, and subsequently, derive the logic specifying dirty data in a specific dataset. Hence, the algorithm devised for data cleansing would learn from examples of dirty data and not rely upon users to specify rules or logic of dirty data.

BayesWipe

Another recent method to improve data cleansing for Big Data involves the application of bayesian networks and is termed 'BayesWipe' (Sushovan et al, 2014). The authors emphasize that traditional data cleansing techniques such as outlier detection, noise removal, entity resolution and imputation cannot provide effective solutions in the context of Big Data. The fact that techniques such as *CFDs depend upon clean external reference sets to learn data quality*

rules is one of the major drawbacks in devising effective data cleansing solutions for Big Data. Even devising rules from 'dirty' data is not judged to be a solution that is satisfactory enough (Sushovan et al., 2014). Sushovan et al. (2014) posit that a statistical process underlies the generation of both clean and dirty data; thus, the data source model and error model are used to detect and repair dirty data. Algorithms are generated from the statistical process and are coupled with updated query rewriting techniques. The fact that BayesWipe could also be applied in an online scenario where only the top-k data portion of the data are considered, and the cleansing process is performed while the data is being retrieved add to its improved applicability in the context of Big Data. Empirical evaluations performed over both synthetic and real datasets tend to show improvements in terms of the amount of data cleansing ratios when BayesWipe is compared with CFDs and Amazon Mechanical Turk, but there is still a very large portion of dirty data not cleansed. For example, the offline version cleans only 40 % of the data in a synthetic car database. Another question about BayesWipe concerns the efficiency of the data source and error models which is the foundation of this method. The evaluation results given denote that those models could be improved to lead to higher data cleansing ratios (Sushovan et al., 2014).

Implications for the current research*: The current research would build on the ideas forwarded through BayesWipe through experiments involving it* and use the knowledge learnt to propose an improved data repair method in the context of health data.

Data X-Ray

All the techniques discussed above apply data cleansing methods on the data itself, but do not attempt to correct the *cause* of DQ issues. As most of those data quality issues are reported to be systematic, thus inherent to the *process of data creation*, it is quite reasonable to find meaningful ways to cure the causes of data quality errors as an efficient data cleaning process (Xiaolan et al., 2015). Diagnosing data quality errors in Big Data environments raises some challenges for traditional methods such as provenance analysis, feature selection and causal analysis (Xiaolan et al., 2015). Those challenges are summarized as follows:

i. **Massive Scale:** the high volume associated with Big Data requires parallel computational algorithms and linear time complexity; unfortunately, current feature selection methods are not easy to implement in shared nothing architectures to facilitate implementation of parallel algorithms.

ii. **System complexity:** Data sources for Big Data tools are often various and numerous, causing provenance analysis to be impractical to carry out as there is often no direct access to the sources of the data.

iii. **High error rates:** Some applications of Big Data such as web-based data might suffer from error rates as high as 70%. This makes causal analysis to be inapplicable with Big Data as causal analysis is based on the notion that errors are rare in a dataset.

Data X-Ray proposes to overcome the above challenges by (1) finding a hierarchical structure of features which best represent erroneous elements with the aim of understanding most important DQ issues, (2) using Bayesian analysis to estimate the causal likelihood of features being associated with potential causes of errors and diagnose those causes using conciseness, specificity and consistency DQDs.

***Implications for the current research**:* This current research study *does not attempt to solve data quality issues at the source of data production*, as discussed by Xiaolan et al. (2015); DQ issues at the source of production could be due to bad organisational processes or other very specific organisational issue such as improper employee training. However, as the DQ strategy followed by this research study is data-driven, recommendations to prevent data quality problems at the source of production go beyond the scope of this work. Furthermore, the 'variety' and 'velocity' properties of Big Data would be a challenge for correcting the source of data quality errors.

Potter's wheel

This technique was designed to operate as data transformation and integration algorithm or solution for data warehouses (Raman & Hellerstein, 2001). At that time, data cleansing solutions were suffering from two main issues; (1) lack of interactivity which led to cleansing procedures to be carried out in batch mode and (2) need for intense user effort, causing data transformation to become a tedious and user dependent process. Potter's wheel brings interactivity to the data cleansing process as users can see the transformation process in real time and without needing to use complex regular expressions. It provides an 'MS Excel' like interface with satisfactory GUI which makes the transformation process less tedious for users. A very interesting feature as part of Potter Wheel's architecture is the 'online re-orderer' module which allows the user to view continuous errors being corrected from specific datasets; this feature could definitely be very useful in the current research due to the velocity aspect of Big Data. Another useful feature is the

discrepancy detector which automatically identifies faulty data based upon transformation rules given by the user.

Implications for the current research: Potter Wheel's suffer from certain deficiencies in the context of Big Data. Firstly, its source data connection only is of ODBC type, which limits its connection mostly to relational databases. Secondly, it relies heavily upon users to ascertain the type of transforms to be undertaken. Finally, it requires a heavy consumption of primary memory. Thus, even if some of its characteristics such as the 're-orderer' idea could be transposed to a data repair tool for Big Data, Potter's Wheel does not seem highly adequate as data cleansing algorithm for Big Data.

NADEEF

The lack of end-to-end off-the-shelf automated solution to automate data error detection and data repairs prompted the design of this system. Heterogeneity for allowing users to express rules in different methods and being able to act upon those different methods are its main strengths. Thus, users can express rules easily and are abstracted of the data detection and reparation processes (Amr et al., 2013). In this way, the human intervention part of NADEEF resides in focusing on domain expertise for rules expression, while the software caters automatically for reparation, resulting in an improved performance so important for Big Data context. However, a method like 'BayesWipe' seems to be more advantageous for Big Data as the rules' specification process is more automated, but the velocity aspect of Big Data might be a challenge for algorithms to automatically learn from dirty datasets. NADEEF was also evaluated with some health-related datasets, and preliminary results after data cleansing denote more than BayesWipe's 40% ratio of clean data. However, there is a lack of information about whether the datasets used for the design experiment of NADEEF exhibited Big Data properties.

Implications for the current research: The current research stands to benefit from a fusion of ideas coming from NADEEF, BigDansing and BayesWipe.

Febrl

Febrl (Freely extensible biomedical record linkage) is primarily a data matching tool written in python with the source code made freely available from https://sourceforge.net/projects/febrl/. Apart from record linkages, it also contains techniques for data deduplication and data cleansing

(Christen, 2008). Its main aim is to allow researchers and practitioners to experiment with a robust record linkage tool which is easy to use in the health/medical field.

There are usually three distinct set of activities which could be performed with Febrl; (1) cleaning and standardisation of a dataset, (2) deduplication of a dataset and (3) linkage of two datasets. Febrl contains predefined standardisers which could be applied for fields such as 'address' and it makes use of a combination of rule-based approach and probabilistic hidden markov model. The comparison module of Febrl contains around 26 similarity functions for string, numerical and date/time comparisons which would be used for data repairs. Each compared record pair will be assigned a weight vector based upon both supervised and unsupervised classification techniques. For supervised technique, a support vector machine implementation is customized whereas 'KMeans' is an example of an unsupervised technique; both would cluster record pairs into match and non-matching groups. Finally, the 'TwoStep' classifier is an unsupervised approach which selects highly probable matches and non matches of records, but also builds a training set for a binary classifier.

***Implications for the current research**:* The fact that Febrl uses machine learning for record linking and its free availability makes it an ideal tool to consider while developing the different algorithms which this research seeks to attain. However, the main constraint of Febrl is that it focuses upon specific types of data cleansing activities. Some further pitfalls relative to the current work is that Febrl focuses upon biomedical data and that it depends upon pre-Big Data era technology; this tool could be benchmarked in the experiments for data repairs of this current study, depending upon whether biomedical data is used as benchmark dataset for the experiments.

2.2.10 Data quality rules

Enforcing Data quality rules (DQRs) are integral activities for DQ issues detection in traditional datasets. Prior research in the field show that DQRs are being enforced via various methods such as Functional dependencies (FD), Conditional Functional Dependencies (CFDs), Dedupalog, Integrated Constraints (ICs) and Bayesian networks amongst others (Chu et al., 2014; Yakout et al., 2010; Yeh & Puri, 2010). All of those techniques have been evaluated through several researches with the common purpose of improving the efficiency of the data cleansing or repairing activities; the efficiency being quantified as both the amount of time to perform cleansing of datasets and amount of errors being corrected in given datasets. Another common

theme through all those research studies is the fact that rules are discovered out of the characteristics present in the data. One study focusing on EHRs systems found that there could be around 11000 data quality rules applicable (Huser et al., 2018). This indicates the vast amount of data quality rules which could be applicable in just one domain and therefore it could be quite unrealistic to rely upon human experts to derive all the rules to ensure optimum data quality. Other algorithms such as Raha/Baran (Mahdavi et al., 2019) also produce DQRs automatically from the data itself, but this process is extremely computationally extensive and therefore might be prohibitive in the context of Big Data.

Yeh and Puri (2010) aimed at increasing consistency in datasets by discovering rules for more efficient CFDs. Challenges with increasing consistency according to Yeh and Puri (2010) are (1) it is a labour-intensive process and (2) rule discovery is largely a manual process which relies heavily upon subject matter experts. Furthermore, methods for discovering CFDs have been reported to have difficulties to scale for relations having a large number of attributes and they are not robust with datasets having a high level of dirty data. Yeh and Puri (2010) carried out research work aimed at increasing consistency in datasets by discovering rules for more efficient CFDs. They developed an approach called 'CFinder' which follows the following main steps (in Figure 2.2 below) to *automatically* generate better CFDs:

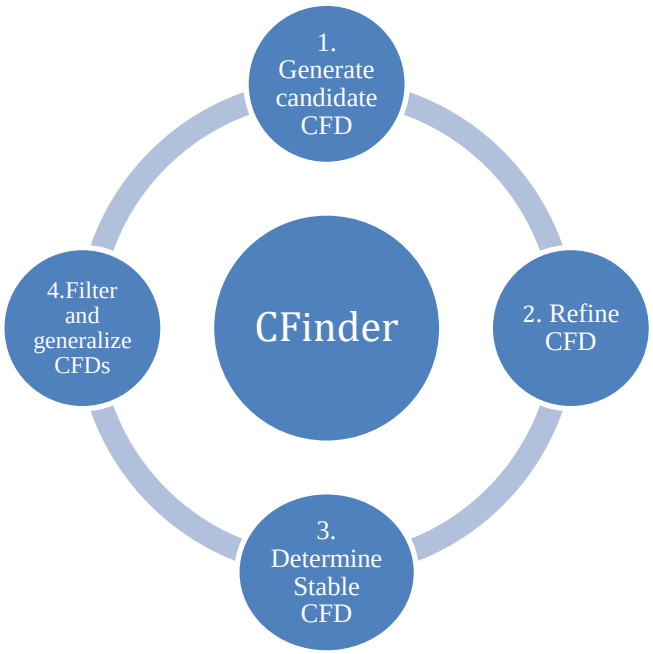

Figure 2.2: Main steps of CFinder (Yeh and Puri, 2010)

Even if CFinder outperforms CFD-TANE in terms of the recall and precision metrics, Yeh and Puri (2010) investigated ways to improve CFinder with (1) the use of heuristics to improve scalability (2) application of industry ontologies to determine which attributes are related for the pruning process and (3) exploration of other metrics to eliminate weak CFDs.

There could be pitfalls relying upon automated data repair solutions based on DQRs, especially in use cases dealing with critical data. However, the involvement of human users to validate data repairs means the response time of data quality tools degrade considerably. Thus, an interactive method which performs some proportion of automatic repairs while allowing users to validate repairs was proposed (Yakout et al., 2010). The method involves generating repairs to only *the top-k most important violated rules* and ranking the most beneficial repairs from the users' perspectives for their validation (Yakout et al., 2010).

Implications for the current research*:* Experiments comparing the rule ranking method with other techniques such as the *Greedy* and *Random* algorithms tend to demonstrate that the rule ranking method is more efficient. However, in the context of Big Data, there are several challenges which might be addressed by the current research study:

1. How practical is it to have the user validating repairs, even for the top-k rules, when there could be the assumption of a huge size of top-k repairs to be undertaken for Big Data?
2. There should be some measures for the computational complexity of the rule ranking algorithm as it involves nested loops and user interactions. Furthermore, the stopping condition of one of the loops equals to the fact that there is no more dirty tuples in a given dataset. In a Big Data scenario, with the high velocity of data production, this could well result in the algorithm generating infinite loops!!
3. The top-k repairs would invariably be linked with the use case for which analytics are being applied in a Big Data dataset. As already questioned before in this research, there is a legitimate question whether to transform the data repairs according to one particular use case and thus update the original dataset OR create a copy of the corrected data while keeping the original dataset for other use cases.

RULEMINER is another system to discover DQRs which aims to address the main limitations of existing rules discovery methods which have been reported to be (Chu et al., 2014):

- Existing rule discovery algorithms are usually designed for a single rule language, thus unable to discover many useful rules for a dataset.
- Most existing DQR algorithms generate many rules, where many of those rules are not adequate.
- Manual evaluation of the output of rules is a time-consuming process.

RULEMINER discovers rules expressed as Denial Constraints (DCs) which is supposed to subsume FDs and CFDs. However, it is quite unclear whether DCs would subsume more elaborate rules such as for semantic interpretation of data as discussed in the data cleansing section (Hima, et al., 2011) of this current research and thus, the first limitation listed above is unsure to be addressed by RULEMINER. Another issue with RULEMINER is the fact that there is a dependence upon users to validate repairs in terms of its Negative Example-Positive Examples pairing; with Big Data, this could result in repairs that are too computationally costly. However, this method seems to be a very user friendly with a front-end interface which allows the user to specify the maximum number of errors to display for a given discovered rule (similar to the top-k notion) and a filtering option allows the user to focus upon certain rules depending on a given use case.

2.2.11 Data Quality frameworks/methodologies

Data cleansing is a very important step as part of managing data quality for a proposed data quality methodological approach. This step is further decomposed into four ways to clean data namely (1) *correcting* defective data elements, (2) *filtering* which involves removing bad data, (3) *detecting and reporting* when it is not cost effective to correct bad data and (4) *preventing* which involves avoiding the causes of producing data of bad quality (Abdullah et al., 2015).

A two-way approach to DQ consisting of (1) being data driven and (2) being process driven, was proposed by Taleb et al., (2018). The data driven component is made up of steps such as data cleansing, filtering and approximation while the process driven component involves catering for processing and analytical activities. The results show that combining the two components lead to improved quality enforcement. The data driven component makes use of pre- and post-Big Data quality evaluation which applies metrics for data accuracy, completeness and consistency DQDs to measure improvement after data repairs done. Pre-processing quality evaluation refer to the evaluation of the data before the processing and analytics stages. It makes use of metrics

relative to accuracy, throughput and response time. These metrics are also applied in the last evaluation stage, namely after the processing and analytics stage.

***Impact upon the current research*:** This current research focuses on data driven initiatives and might adopt the pre- and post-Big Data quality evaluations. However, the way that the data repairs are applied might be different, as the statistical methods aiming to resolve mostly data incompleteness is considered not to be adequate for different data types possible with Big Data.

A data quality framework consisting of a number of interesting steps was proposed (Taleb et al., 2015) and described in the Figure 2.3 below. This framework is referred as BDQPF.

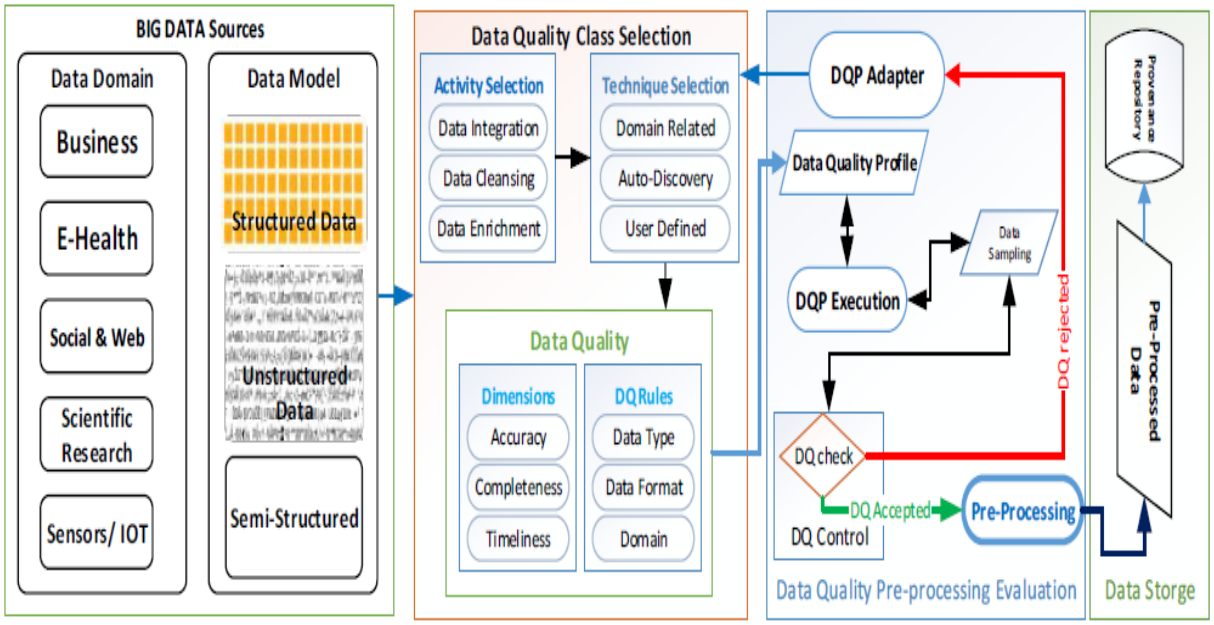

Figure 2.3: Steps of Big Data quality framework (Taleb et al., 2015)

Data quality profiles are created for each data source based upon DQ class selection. Actual DQ pre-processing is based on the activities forming the 'activity selection' substage of 'DQ Class Selection' stage. The concerned activities are namely data integration, data cleansing and data enrichment. These data processing activities are guided by rules which are related to the domain of data use, discovered partly from the data itself (auto-discovery) and partly from the user.

***Impact upon current research*:** The framework in Figure 2.3 possesses many similar ideas and characteristics to what is proposed in the current research study. Data from a specific domain is collected, and exact DQ requirements are ascertained in terms of DQDs and data types. The major

difference is that the current research study proposes exploring the use of ML techniques to perform the specific data improvement activities, and thus does not consider data integration and data enrichment. With the use of ML techniques, there is no need to create user defined DQ rules and to maintain DQ profile list also. Furthermore, some techniques which have been used to represent a Big Data scenario could be replicated in the current research study.

There is another Big Data quality framework which is very similar to the one proposed by Taleb et al. (2015) by Juneja and Das (2019). The framework is illustrated by Figure 2.4 below:

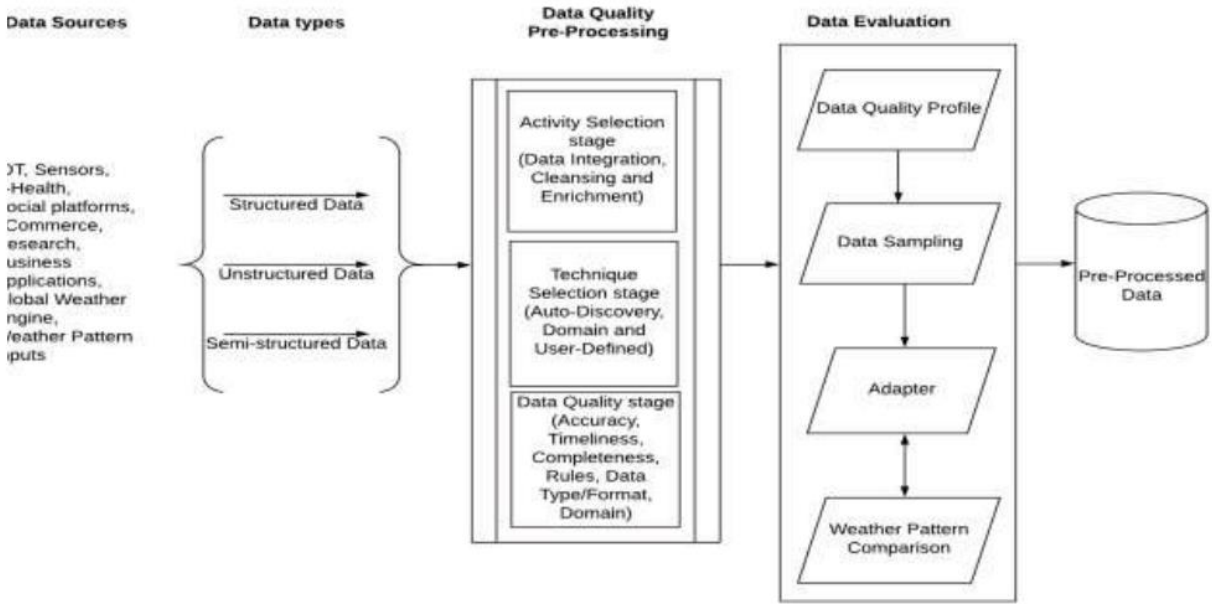

Figure 2.4: Big Data Pre-processing framework (Juneja & Das, 2019)

Upon comparison of this framework with BDQPF, many of the stages were found to have already been discussed. For some of the differences noted with BDQPF such as the data evaluation stage, there lacked details precisely describing the steps such as what is exactly involved in the 'Adapter' step and the implication upon Big Data quality. However, the fact that different authors discussed very relatable stages pointed to some pattern of steps which might be expected from a potential data quality methodological approach.

A very interesting Data Quality framework known as DQF4CT was proposed for general data mining tasks (Corrales et al., 2018). As part of DQF4CT, a conceptual framework for data cleansing tasks is given as below:

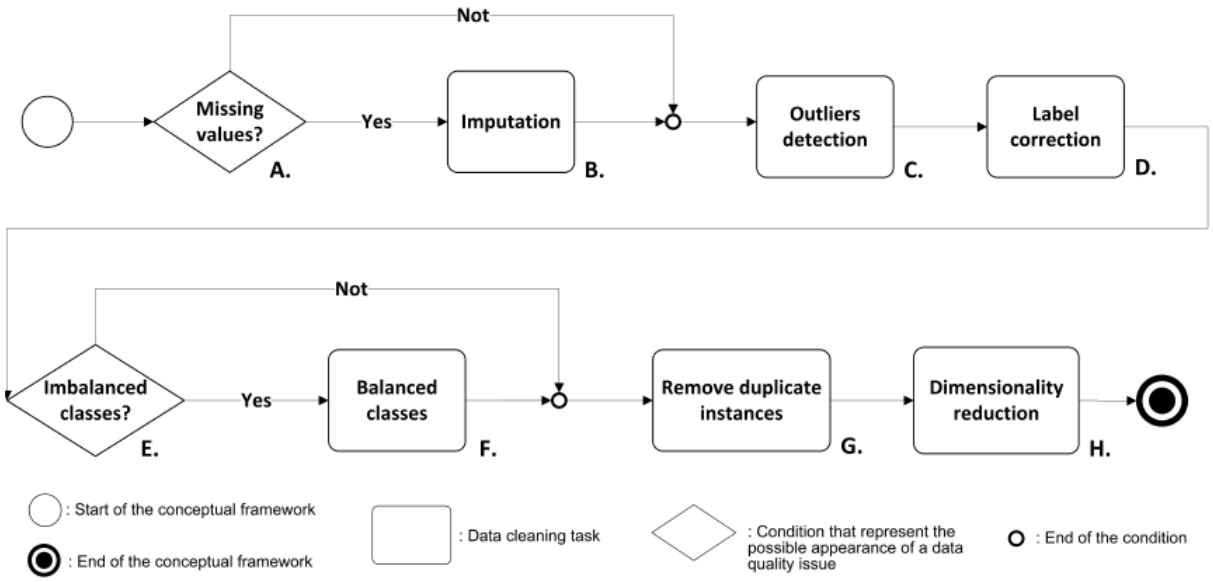

Figure 2.5: DQF4CT (Corrales et al., 2018)

The main steps are being provided by alphabets A to H in Figure 2.5 above. They consist of important data cleansing activities such as checking and dealing with missing values, dealing with outliers and dimensionality reduction amongst others. The main purpose of DQF4CT is to provide a framework to improve DQ for data mining tasks, based upon weaknesses found after surveying current data mining tools. Hence, it includes purely classification based DQ improvement steps such as balancing classes and label corrections. However, these classification tasks are not specifically for Big Data, even if during the evaluation of their work, the authors did make use of very large datasets with high dimensions. Even if application of DQF4CT did improve data mining results on datasets concerned, the authors acknowledged that domain knowledge is important when applying data quality frameworks.

Impact for current research: Many steps proposed as part of this data cleansing framework as explained above would benefit from further investigations, for example, the use of machine learning to support and facilitate certain of the above steps such as dealing with missing values and outlier detections. Certain methods could be replicated to carry out the evaluation of the proposed data quality methodological approach.

2.3 Conclusion

The aim of this chapter is to consolidate knowledge pertaining to connected ideas and concepts which can inform the current research study. The first section explored generic ideas such as Big Data technologies, types of health data and machine learning, amongst others. This was to inform readers about the role of certain of these generic ideas, and the impact they might have upon the current research study. Through this section, the breadth of the current research study can be better appreciated, as it covers several fields and domains.

The second section explored ideas which are more connected with the main research question and different research objectives, as discussed in Chapter 1. The ideas in the different research studies were analysed from the prism of the current research study, such that critical annotations or explanations of each section are discussed. Thus, the importance of each idea being discussed can become more apparent, and better inform the current research study.

Research gaps were exposed and explained throughout this chapter. These gaps will guide future activities such as the general research methods which might be applied, filtering out ideas and concepts as part of research to inform the different stages of the proposed methodological approach and obviously, looking out for potential answers to them.

As this research spanned over several years, this chapter has been constantly amended and cross checked. As and when relevant ideas were discovered, they were integrated in this chapter. Finally, this chapter contains relevant information about the current situation for this current multi-disciplinary research study.

Chapter 3: Methodology

3.1 Introduction

The goal of this chapter is to describe the research design and the different tools and techniques that have been applied to answer the principal research question and inform the different research objectives. It also aims to evidence the main components that form part of a proposed methodological approach to optimize data quality for EHR Big Data. This chapter also introduces the main datasets used in conjunction with the different experiments carried out to detect specific DQ issues and data repairs. The characteristics and intended method of use of each benchmark datasets considered for experiments would be probed and explained to justify degree of adequacy to fulfil the research objectives and inform a potential methodological approach.

3.2 Research philosophy, approaches and techniques

In terms of **research philosophy,** this research follows positivistic principles since one very important part of the proposed data quality methodological approach aims at *evaluating* new data quality detection and data repairs algorithms. Furthermore, another research objective informing the methodological approach is to identify which DQDs are most important for Big Data use in the health industry. Hence, identification and evaluation are the main research activities involved and they are linked with positivistic approaches. The phenomenological philosophy is simply ruled out since the research is not linked with human participants such as Big Data or DQ experts and as such does not directly influence human behaviour.

The nature of this research could be termed as innovative due to the small number of existing research studies in Big Data quality and the few existing tools available to enable Big Data quality, specifically for the health sector. Thus, the current research could also be described as **exploratory**. The fact that this research study applies some ideas such as the investigation of unsupervised machine learning techniques for detection of dirty data within a well specified Big Data source reinforces the notion of exploratory research. However, since one of the first research objectives is to identify the most important DQDs, this gives the research a descriptive aspect. Hence the research can be classified as mostly exploratory, with some descriptive elements integrated within it.

Concerning **research approaches**, the scope of this study encompasses many different approaches which could be explained as follows:

- A *quantitative aspect* as there is a subsequent element of comparison and evaluation of algorithms (DQ issues detection and data repairs) to determine their effectiveness in the context of Big Data for the health industry. Furthermore, the research design is well detailed and structured, and those two properties adhere to the categorization of quantitative research.
- A *qualitative* aspect with regards to the selection and interpretation of ideas coming from previous research studies such as to posit the most important DQDs.
- Since the research involves real world data as part of the benchmark dataset used for experiments, the research approach could also be termed as *applied research*.
- A *deductive approach* is also used. By experimenting with several machine learning algorithms and evaluating them, the research study pinpointed which AI/ML algorithm is most efficient to segregate dirty data from clean data.

Therefore, this exploratory type of research applied both qualitative and quantitative approaches. The qualitative aspects are detailed in parts (1) and (2) below and entail analysing content from existing research studies to deduce the DQDs and ML algorithms most relevant for Big Data use in the health industry. The quantitative aspect is detailed in part (3) below and involved the use of controlled experiments to determine whether data repair algorithms developed during this research study are more adequate than existing data repair algorithms and solutions relative to DQDs identified in part (1) below.

The following details the main steps which are part of the research methods used:

(1) The inner hermeneutic cycle (IHC) (Weiskopf & Chunhua, 2013) method was adopted as it was well adapted to uncovering ideas from existing literature in emergent domains; basically, it consisted of reading, searching, sorting, selecting, acquiring, identifying, and refining content from previous research connected to DQDs. The original intention was to limit the IHC with DQDs for Big Data specifically, but due to the novelty of this research area and few corresponding research, the scope was widened to accept general DQDs. The search also originally intended to include only peer reviewed journals with search query such as (“data quality dimensions” OR “information quality dimensions” OR “Big Data quality dimensions”) AND (“health datasets” OR “health databases” OR “Electronic health records”). Then, for each journal retrieved, an abstraction of the most

important DQDs was derived. However, as detailed below, the search also allowed some other sources of information which was judged to be relevant with the knowledge of DQDs.

(2) Latent Semantic Analysis (LSA) (a statistical analysis of word to document similarity method) was used as an additional method to minimize the effects of author-based subjectivity of the IHC approach (Kulkarni et al., 2014; Casakin and Singh, 2019). The LSA is a widely accepted technique to determine the importance of terms found within sets or corpus of documents. Highlighting most important terms in an area is different from the fields of topic modelling, document modelling and highlighting thematic patterns. The abstracts coming from the same set of documents coming from existing research studies involved in the IHC was used and fed into an LSA algorithm developed using Python 2.7 and specialized libraries. LSA infer the importance of terms per a whole set, often known as corpus, of documents by applying a cosine similarity index. Therefore, this method is more adequate compared to information retrieval which would entail simple computations such as word counts to denote importance of terms.

(3) . A comparison and evaluation of some of the most relevant AI/machine learning algorithms was performed in the context of DQDs identified in (1). The most appropriate evaluation measures were chosen and used depending upon the algorithms implemented and also on insights from data present in the real-world dataset chosen.

(4) Partly automated data repair prototype was developed for Big Data involving less computational complexity. It was compared with state-of-the-art data cleansing tools such as Raha/Baran, ActiveClean, KATARA and NADEEF amongst others. The DQDs discovered in (1) was used to determine whether a new proposed data repair prototype performs better than existing ones. E.g, if ‘accuracy’ is one of the discovered dimensions, metrics and measurements used to measure ‘accuracy’ for data quality are applied. The AI/ML algorithm identified in (3) is then used to identify data which would need to be repaired by the proposed data repair method. It must be noted that most algorithms investigated perform both dirty data detection and data repairs, but few of them focus upon outlier data repairs as what this thesis focuses upon.

3.3 Inner hermeneutic cycle

As the domain of the current research study is very innovative, there exist few directly corresponding existing research studies. The first research question relates to ascertaining which DQDs would be the most important for Big Data in the health industry. Even if there are numerous existing research studies discussing DQDs in general, with some in the context of the health industry, there is no existing knowledge which would forcibly posit the most important DQDs of Big Data in the health industry. This research applied a classification of the structural properties of data quality based on the interpretation of DQDs discussed in a mutually exclusive way (Weiskopf & Chunhua, 2013). The choice of the DQDs to be interpreted was related to the structure of data discussed as part of existing research studies.

Hermeneutics could be loosely described as the analysis and interpretation of texts and literature. The inner hermeneutic cycle (IHC) consists of searching, sorting, selecting, acquiring and reading, identifying and refining ideas within existing literature. Since the discovery of DQDs most important to big health datasets is of a qualitative nature, potential research methods are the use of interviews of data quality managers/experts or the integrative review of existing literature using the IHC. The interview method, however, was not used in this current study for the following reasons: the almost non-existence of Big Data quality managers/experts in Mauritius; difficulty to get into contact with data quality managers that were external to Mauritius and most importantly the fact that many data quality managers in the health industry have not yet adopted a proper framework in the context of Big Data.

On the other hand, IHC or similar integrative reviews of literature are research methods which have already been applied in the context of health and data (Weiskopf & Chunhua, 2013) and in the context of Big Data quality (Batini et al., 2016). Batini et al. (2016) quote other authors (such as Beyea & Nicoll, 1998; Torraco, 2005; Whittemore & Knafl, 2005) who have also made use of IHC. This method is suitable for any domain of research which is emergent in nature such as where there are very few real life or practical applications of concepts discussed. This in-depth investigation of existing literature allows the formulation of theoretical frameworks which could be validated by further practical experiments once the emergent technology or domain area becomes more ‘main-stream’ and adopted by industry players.

Therefore, despite the existence of very few academic publications related directly to data quality for Big Data in the health industry, an exploratory approach of data quality reported in general for Big Data was chosen. The search activities were carried out on research databases of different areas such as IEEExplore, ACM, health.gov, SCOPUS and MEDLINE since they were the most well-known ones. The keywords and search operators used were as follows: data quality in big health datasets, data quality AND Big Data, data quality dimensions AND health datasets, information quality AND health datasets, very large datasets AND data quality. Regarding the sorting phase, only recent articles were taken, ranging from 2006 up to 2016. This was later refined from 2006 to 2021 and the results were updated. Other criteria for selection were: (1) the popularity and hence acceptance of the article determined by the number of citations obtained wherever possible; (2) the interpretation of ideas or concepts put forward by authors and their relevance related to this current work. With regards to the latter, some research studies which involve data quality with machine related or sensor-based data could have initially been thought quite irrelevant for this research, but after analysis of some of those research studies, corresponding ideas in terms of similar types of data to health industry could be denoted. Hence, those research studies were included for further decoding and analysis. There was also the need to establish what types of data could be considered more representative of use in the health industry and this was an additional factor in the rationale for scoping the boundaries of the current research study. More in-depth details about this section of the research study are discussed in chapter 4.

3.4 Statistical techniques to infer importance of terms

There are different domains such as topic modelling, document modelling and relevance ranking where algorithms are contributing towards working more efficiently with documents and terms. An example of use of topic modelling is to perform sentiment analysis by constructing probabilistic models with methods such as Probabilistic LSA (p-LSA), LSA and Latent Dirichlet Allocation (LDA) (George and Birla, 2018; Harada et al., 2020). Thus, topic modelling is well suited to understand polysemy of words and to automatically classify a topic in text data, where LDA and p-LSA provides more efficiency compared to classic LSA. On the other hand, a well-known example of relevance ranking is search engine optimisation which can be implemented with deep learning algorithms such as DeepRank (Pang et al., 2017). However, the aim of this doctoral study's first research objective is **different from both topic modelling and relevance ranking** as the goal is to infer the importance of different terms within a small corpus of

documents. Topic modelling is based upon the number of frequency of terms in a corpus. The logic for this current doctoral study is that the frequency of a particular term does not provide a good estimate of the importance of the term. Concretely, for example, if the DQD 'timeliness' is mentioned 100 times throughout a corpus this should have nothing to do with its importance in the area, as the use of 'timeliness' might have been used in many other contexts different from DQDs. Hence, term to document similarity measure is critical and the LSA application with the implementation of cosine similarity provides this (Landauer et al., 1998). Furthermore, the goal of using an algorithm for this section of the research is to provide objective results of discovery which do not suffer from potential subjectivity associable with integrative review method. Hence, given the small size of the corpus involved and the clearly specific aim of finding importance of terms to documents for this section of the study, the LSA is preferred over methods such as p-LSA, LDA and DeepRank.

LSA is a statistical method for estimating the meaning of terms based on linear combinations of underlying concepts. It had been applied in a variety of fields ranging from operations research in management, library indexing improvement, and search engine query performance optimisation to chatters perception on social networks (Kulkarni et al., 2014). The fact is that wherever the importance of terms needed to be extracted from a set of text data, LSA is a technique worth considering. LSA is a technique created some decades ago, in 1988. Decision makers want to have the ability to work with data grouped as a corpus, but the semantics or choice of terms used might be different according to different authors of documents and therefore, there should be ways to create inductive relationship between terms and documents. The LSA is the most highly rated algorithm which provides this inductive relationship relative to LDA and p-LSA (George and Birla, 2018).

3.4.1 Application of LSA for importance of DQDs

Latent Semantic Analysis was used to determine the meaning of words and passages of large text corpora (Landauer et al., 1998). LSA applies single value decomposition (SVD) matrix, which is a mathematical decomposition technique very similar to factor analysis and which is largely recommended for text analysis (Landauer et al., 1998; Harada et al., 2020). SVD reduces dimensional representational of a text matrix of words to documents, whereas the application of cosine similarity provides the importance of a given word for a corpus of documents. With cosine similarity measures, the importance of a word in a context might be greater compared only to the count of word in the context, because cosine similarity would forecast the importance of the

particular word in a projected infinite number of articles. Alternatively, even if a word appears frequently in a particular research paper, the application of cosine similarity might conclude in a low importance for the particular word if the factor analysis algorithm predicts that this high occurrence is only for this specific research paper and might not hold for the research domain area holistically.

Hence, with the application of LSA, this research aimed to sort the importance of the forty-six (46) DQDs uncovered via the application of the inner hermeneutic cycle upon existing research studies. LSA evaluated the similarity occurrence of those individual 46 DQDs per previous research studies by the application of a cosine similarity measure. More specifically, this current research study applied the word to passage relations via the application of semantic similarity. This resulted in a tabulation of research studies and their associated most important data quality dimension term used.

3.5 Experiments upon EHR Big Data

It is possible to get access to free datasets related to the health industry holding different types of data in formats such as HTML, RDF, CSV, JSON and XML. Examples of such datasets may be accessed through websites such as www.healthdata.gov, www.data.medicare.gov, www.datasciencecentral.com amongst others However, many of those datasets are already fragmented into several distinct components. For example, from www.data.medicare.gov, there is the possibility of downloading different categories of datasets such as 'hospital compare data', 'physician compare data' and 'supplier directory data'. The 'hospital compare data' dataset is itself made up of over 50 different files.

One important task for the research study was to select and choose which datasets could most appropriately serve the experimental purposes designed for DQ issues detection and data repair algorithms comparison and evaluation. Most existing research studies concerned experiments performed upon Big Data make use of 'CSV' based datasets. However, some more recent studies do make use of tools such as Hadoop or Apache Storm as data sources representing Big Data.. These technologies are implemented over cloud-based services, and the current study has simulated Big Data operations over EHR and BigQuery datasets by making use of Google Cloud Platform (GCP). Therefore, aligning with the norm of using 'CSV' based dataset, this current study implemented datasets populated with real-world health data over the GCP and benefitted from increasing processing power and availability of more RAM compared to a local desktop machine which allowed to work with bigger sets of data. Furthermore, experiments involving some algorithms were carried out via the use of BigQuery ML over a publicly available EHR BigQuery repository.

Having obtained satisfactory benchmark datasets for the experiments, it was very important to have a precise set of statistics pertaining to the ratio of dirty to clean data present according to each DQD identified. Metrics and measurements discussed in the literature review section were applied on the different datasets in the initial stages of the experiments.. For this research study, statistical functions were used as data exploratory techniques.. This knowledge of the amount of dirty data is essential to compare and evaluate the machine learning algorithms applied upon the chosen real world benchmark datasets.

3.5.1 Detection of DQ issues

For a better understanding of the second main research objective, it was imperative to use the baseline knowledge of DQDs and exact manifestation of DQ issues present as part of the selected real-world benchmark datasets. Carefully selected machine learning algorithms were applied on the benchmark datasets and evaluation measures based upon metrics per DQD identified as most important for EHR Big Data as per Chapter 4 were carried out. The exact algorithms to be implemented and compared was extracted from knowledge retrieved in existing research studies connected with data imputation and detection of data outliers. As this research study made an assumption that unsupervised learning method not requiring annotated class labels could be more adequate for detecting DQ issues for Big Data, the experiments included at least one of them.

As mentioned in the last section, the purposes of comparisons and evaluations were two-fold; (1) to confirm or infirm a hypothesis claiming that unsupervised methods should detect DQ issues better according to predefined DQDs (2) to use clear metrics and evaluation techniques to rank ML algorithms. It must be noted that the experiments were carried both on local desktop computing and cloud-based architecture. The comparisons enabled the researcher to state which algorithm is better for data imputation and detection of data outliers for EHR Big Data.

3.5.2 Experiments for data repairs

The final research objective aimed to produce a better data repair algorithm compared to some known algorithms as discussed in Chapter 2 (literature review) and Chapter 6. This enhanced data repair algorithm should be expressed in terms of less computational complexity and optimised data correction metrics to minimize insertion on new errors after the data repair process. Consequently, the first step towards this phase involved the development of data repair prototype having those two above mentioned properties. The literature review revealed that there are algorithms which already possess those techniques, albeit not in combination in the same algorithm.

Hence, taking the latest data cleansing algorithms such as BigDansing as a basis, a new enhanced prototype was developed. The next step was to evaluate and measure the performance of the new prototype compared to existing algorithms and existing software solutions based upon firstly computational complexity of each algorithm, secondly the proportion of user intervention in

terms of user actions needed during the data repair process and thirdly, the amount of correct data repairs done .

All those experiments were carried out in a controlled environment and in a manner of execution to eliminate/minimize the impact and consequence of uncontrolled events upon the experiments. Due to the nature of this research involving well constrained real-world datasets and algorithms, the probability of impact of uncontrolled events was almost inexistent. Yet, all due measures are taken so that the experiments and their results remained valid and scientifically acceptable. More in-depth discussion about experiments implemented are detailed as part of Chapters 5 and 6 of this research study.

3.6 Evaluation methods

The proposed methodological approach to optimise data quality for EHR Big Data is made up of different distinct stages and steps. During the investigation to determine the exact nature of these steps, there needed to be clear methods used to evaluate different steps to be able to understand which steps are most efficient and under which conditions. This section details the expected evaluation methods used for the main research objectives of this current research study.

To evaluate the most important DQDs for Big Data in the health industry, the results of both the IHC and LSA were compared against each other. The number of DQDs to be considered could not be ascertained at this stage, but these DQDs should denote a significant importance in the domain of this research study. Upon the comparison of both methods, only common DQDs were considered as most important. More details are provided in Chapter 4.

The experiments used to determine the most relevant AI/ML algorithms to detect DQ issues were based upon specific DQDs. Traditional evaluation criteria used for assessing ML algorithms firstly target the accuracy of predictions made through benchmarks such as precision, recall, harmonic mean, ROC and AUC. All these evaluation benchmarks can be used, but a point of concern was the fact that most of these methods require knowledge of data ‘truth’ sample of known correct and incorrect data items before running the experiments. Thus, plausibility was used as a more suitable evaluation technique for ML algorithms involved with real-world datasets in the absence of ‘truth’ samples. The time taken for executing the different AI/ML algorithms was another criterion used for the ML algorithms evaluations.

The last experiments concerned evaluating data repair algorithms. The same challenges for evaluation cited above were applicable since the same real-world dataset was used and the lack of ‘truth’ samples of correct/incorrect values prevented the application of evaluation techniques such as precision and recall. When applying some datasets used (as part of this thesis) upon other algorithms, classification metrics such as precision/recall and f1 score were used and allowed a certain degree of comparison between some data repair algorithms. Furthermore, mean absolute value was also used to evaluate the imputation algorithm recommended by the proposed prototype over and above the use of plausibility metrics. Computational complexity was assessed using a time measuring function, whereas degree of user involvement required a subjective evaluation in terms of the role of the user during the data repairs and the degree of involvement of the users during the data repairs process.

The main investigations concerning the stages of the proposed methodological approach were themselves research topics. This entailed that more precise literature relative to data quality methodological approaches were consulted and critically analysed. The discovered data quality methodological approaches were assessed in the context of Big Data and the results discussed with the goal of answering the main research question of this study.

Chapter 4: Investigating the most important Data Quality Dimensions for Big Data in health industry

4.1 Introduction

This chapter details the steps pertaining to the findings relative to the first major research objective of this thesis, which is investigating the most important DQDs for Big Data in the health industry. The formats and types of data involved within the health industry are extremely wide ranging from highly structured data (e.g., ICD diagnosis codes) up-to totally unstructured data (e.g., doctor's notes or prescriptions). One of the main research gaps for this study concerned whether the DQDs cited for traditional data use are also relevant for Big Data in the health industry. This chapter addresses this gap. As explained in Chapter 3 of this research study, the research methodology applied to solve this research question is qualitative and is based on integrative reviews of existing research studies. Two methods were applied to increase the relevance and authority of the results from the integrative review. The first method intended to find, extract, sort, analyse and deduce DQDs cited by previous literature is known as the inner hermeneutic cycle (IHC). The second method involved the use of a statistical technique known as latent semantic analysis (LSA) used to determine "word to document similarity" and infer importance of terms per corpus of documents. The use of LSA is mostly as an additional method with the emphasis of bringing objectivity to the integrative review analysis and to prevent the study from suffering from author bias. Eventually, after comparing results of both methods, the most important DQDs were ascertained.

4.2 Integrative review used.

Existing research studies related with this current domain area of investigation of DQDs reveal several different research methods. These approaches are introduced in the section and the reasons why they could not be applied in the current research study explained.

1) Use of surveys: a two-stage survey was carried in a previous research study and resulted in the identification of four categories containing a total of fifteen data quality dimensions (Wang & Strong, 1996). The participants were users of health data; this method presents practical implementation issue as there is a very limited use of health information systems and therefore of health data in Mauritius. Preliminary investigation carried out reveal that

only two sub medical facilities use health information systems in Mauritius, and the health data is primarily used by data entry operators for the admission department. Furthermore, the data system used is far from the Big Data characteristics. Administering remote surveys to international health users of systems was judged impractical and difficult to implement survey validity. Thus, due to all the facts mentioned above, surveys with those data consumers cannot be considered for this research study.

2) Questionnaires: some previous research studies made use of questionnaires being given to web users when trying to assess information quality dimensions for websites (Katerattanakul & Siau, 1999). In this current research study, the focus is not upon web users and hence identification of participants to reply to questionnaires will question both validity and reliability of the work.
3) Another potential research method could have been some expert evaluation through interviewing of data quality managers or data administrators for health information systems. Practical issues abound deterring the application of interview as research method: almost non-existence of data quality managers in Mauritius; difficulty to get into contact with data quality managers which are external to Mauritius and most importantly the fact that many data quality managers in the health industry have not yet adopted a proper framework in the context of Big Data.

Given the practical difficulties to use research methods as explained above, the research study adopted one which is explained as adequate in domains which is emerging and with few practical implementations and users. This method is based upon integrative reviews of existing research studies.

The main steps of the IHC could be described as shown in Figure 4.1 below:

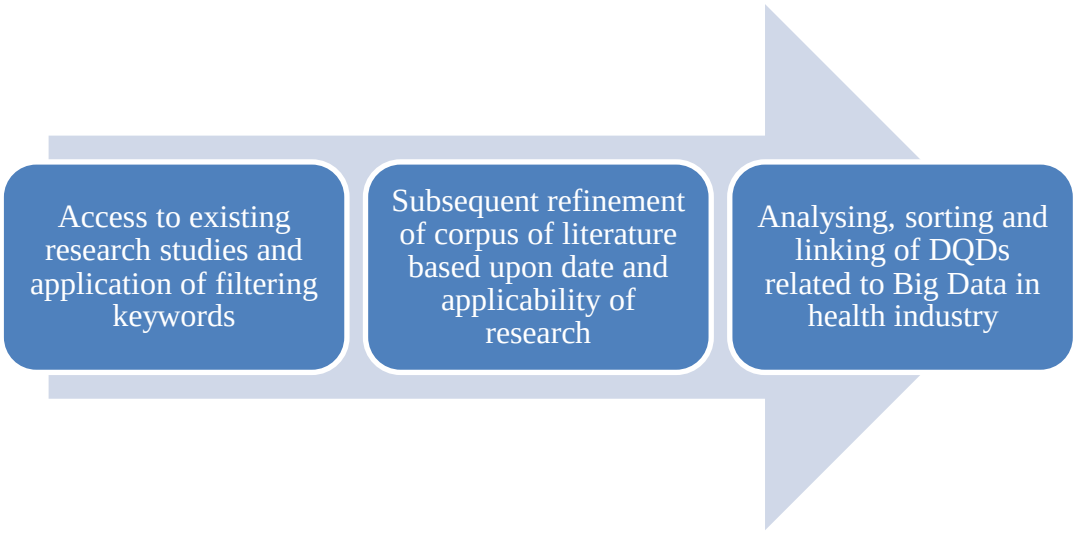

Figure 4.1: Main steps of IHC adopted.

The first two steps are explained in the section below. The analysis, sorting and linking of DQDs stage aimed to develop new knowledge by having a holistic view of secondary sources. Taking the research of Wang and Strong (1996) as a base model, this research study further investigated categories of DQDs which was most closely applicable for the current research. The knowledge of those categories was used to address some of the current research questions associated with the research objective being discussed in this chapter. The aim of Wang and Strong (1996) was to develop a hierarchical framework for organising DQDs. They aimed to associate a set of 15 dimensions into four categories according to the opinion of participants. This resulted to the labelling of the following categories (Table 4.1):

Table 4.1: DQ dimensions categories (Wang and Strong, 1996)

Category	**Description and main DQDs**
Intrinsic	Is explained by data having quality in their own right. (Accuracy, objectivity, believability, and reputation)
Contextual	Highlights the idea that data quality is a factor of the task at hand (value-added, relevancy, timeliness, completeness and appropriate amount of data)
Representational	Includes aspects linked with the format and meaning of data (interpretability, ease of understanding, representational consistency and concise representation)
Accessibility	Emphasizes the role of getting access to data (accessibility and access security)

Considering the explanations of the different categories, the following research four questions for this section were set:

RQ1: Should intrinsic DQ also be applicable for Big Data in the health industry since everywhere there is use of data, those DQDs are supposed to be needed?

RQ2: Would contextual DQ be highly relevant for Big Data in the health industry, as different DQDs could be applicable based upon the context of the use of data?

RQ3: Representational DQ could be less applicable for Big Data in general due to the variety characteristic and the fact that the meaning of data would depend upon the aims of data analysis. Is representational DQ less important for Big Data in the health industry?

RQ4: Accessibility DQ should not be among the most important category for Big Data in the health industry as in terms of access, the datasets are very often publicly available?

4.3 Work undertaken and findings

Based on the IHC methodology outlined in Chapter 3 (Methodology), an initial search was carried out with keywords such as "data quality in big health datasets, data quality AND Big Data, data quality dimensions AND health datasets, information quality AND health datasets, very large datasets AND data quality, data streams and data quality" resulting in thousands of hits. With the SCOPUS database only, there were 2063 matching returns of journals for the query "data quality and Big Data". However, subsequent manual analysis of abstracts of most of the research studies resulted in less than 15 of them mentioning DQDs. For a majority of the 15 research studies, the main aim was not about identifying the most important DQDs, but some DQDs could be interpreted from them. This amount confirmed the emerging nature of the area but also presented a practical issue in terms of having too few related research studies to perform IHC. Therefore, other search queries were devised to be able to get access to relevant existing studies. The term "health data" was applied to search journals and online resources focusing on health informatics.

Manual searching of some journals such as 'Data Science Central' and 'Journal of Data and Information Quality (JDIQ)' revealed that there were some potentially applicable research articles which were not being highlighted by the search criteria mentioned above. For example, with JDIQ, out of around 160 matches with the broad key terms of "data quality", around 15 of them could be linked to either DQDs or data quality in health datasets or data quality with Big

Data but none of these 15 was matched with the search criteria specified above. This could be explained due to different terms used as part of titles of journal papers and the complexity of those titles such as “Challenges in data quality: the influence of data quality assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe”.

The source of some existing research studies also varied in terms of authority; not many sources originated from refereed journals or reviewed conference publications and therefore articles and resources from health and DQ web sites were consulted. For example, the Centre for Disease Control and Prevention (CDC) website generated over 100 matches just for the search criterion of ‘data quality dimensions’. No additional results were returned with most of the search terms mentioned at the beginning of this chapter. With the search term ‘data quality in big health datasets’, there were around 24 results and those which were retained were attributed greater weight for subsequent analysis. Results from some search terms from the CDC website were not research papers but reports and manuals such as ‘National Health and Nutrition Survey Anthropometry procedures manual’. These types of documentation were either not considered or resulted in low weight. Some of the results were pages which contained further links only to abstracts of research papers but without the complete research study. In some cases, the abstract contained enough information relative to DQDs and therefore, the author had to look for the complete paper. However, as CDC is a global centre with enough recognition and authority, the DQDs mentioned in the different reports were considered even if a lesser weight were given to those reports compared to journals or conference papers. The final amount of research papers or articles included for further reviewing amounted to 54.

The 54 papers retained for the IHC were analysed to understand and assign adequate importance to the DQDs discussed by the research studies concerned. An initial organisation of the ideas is summarized in Table 4.2 below. A weight (L: Low, M: Medium, H: High) was assigned taking into consideration the degree of alignment of a research study to the context of evaluation of DQDs for Big Data in health industry. This method of data evaluation was inspired by integrative reviews discussions in nursing where reports were coded on a two-point scale (Whittemore and Knafl, 2005).

Table 4.2: Integrative review results with details of weights

Title of research article used	DQ dimensions	Weight
Discovering dependencies among DQ Dimensions: A validation of instrument (Panahy et al., 2013)	Accuracy, Completeness, Consistency, Timeliness.	L
A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical research (Khan et al.,2012)	Consistency, Completeness, accuracy	M
A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments (Jones et al., 2017)	Accuracy	L
Completeness and accuracy of data transfer of routine maternal health services data in the greater Accra region (Amoakoh-Coleman et al.,2015)	Accuracy, completeness	L
Valid comparisons and decisions based on clinical registers and population-based cohort studies: assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database (Jacke et al.,2012)	Accuracy, completeness	M
Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges (Langley et al., 2006)	Accuracy	L
A Hybrid Approach to Quality Evaluation Across Big Data Value Chain (Serhani et al., 2016)	Accuracy, Completeness, Consistency, Timeliness	H
From Data Quality To Big Data Quality (Batini et al., 2015)	Accuracy, completeness, Accessibility, Trust, Readability, consistency	M
Challenges in data quality: the influence of data quality assessments on data availability and completeness in a VMMC programme in Zimbabwe (Xiao et al., 2016)	Availability, completeness	L
Classifying, measuring and improving the quality of data in trauma registries: A review of the literature (O'Reilly et al., 2016)	Accuracy, capture, completeness	M
Data quality and the electronic health record (EHR) (Maxwell-Downing, D.,2011)	Accuracy, completeness, trust, legibility	L
Pre-charting patient care information (Giarizzo-Wilson, 2011)	Accuracy, completeness	L

Data Challenges in Disease Response: The 2014 Ebola Outbreak and Beyond (Varshney et al., 2015)	accuracy, privacy and security, heterogeneity, provenance and trust, availability, completeness	M
Data Mining Consulting Improve Data Quality (Li et al., 2007)	Completeness, reliability, correctness, Consistency, 'minimality'	L
Data Quality: A Survey of Data Quality Dimensions (Sidi et al., 2012)	Accuracy, Completeness, Consistency	L
Data Quality by Contract – Towards an Architectural View for Data Quality in Health Information Systems (Weber et al., 2015)	Correctness, provenance, currency, plausibility	M
Data Quality Problems When Integrating Genomic Information (Leon et al., 2016)	Accuracy, Completeness, Consistency, currency, reliability, uniqueness	H
Data representation factors and dimensions from the quality function deployment (QFD) perspective (Pinto, M., 2005)	Relevance, Consistency, Accuracy, Currency, comprehensiveness, Format.	L
Defining and Improving Data Quality in Medical Registries: A Literature Review, Case Study, and Generic Framework (Arts et al., 2002)	Accuracy, completeness, clarity, format	M
Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research (Weiskopf & Chunhua, 2013)	completeness, correctness, plausibility concordance, currency	M
Ontology-Based Data Quality Framework For Data Stream Applications (Geisler et al., 2011)	Accuracy, Completeness, Consistency, Timeliness, confidence, data volume	M

Improving Health-Care Statistics Through Electronic Medical Records and Health Information Exchange (Bell, K.,2007)	Accuracy, Completeness, Comprehensiveness, Timeliness	L
A Methodology to Evaluate Important Dimensions of Information Quality in Systems (Todoran et al., 2015)	Accuracy, Completeness, currency, reliability	H
The influence of calibration method and eye physiology on eye tracking data quality (Nystrom et al., 2012)	Accuracy, Precision	L
Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work (Huang et al., 2011)	Accuracy, usefulness, accessibility, relevance, security	H
9th conference on health survey research methods (Aday & Cinamon, 2010)	Accuracy, reliability, consistency	H
Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies (Cure, 2012)	Accuracy, completeness	M
Improving the Predictive Power of Business Performance Measurement Systems by Constructed Data Quality Features? Five Cases (Vattulainen, 2015)	Completeness, redundancy, accuracy, representativeness, consistency	L
Contrasting the Dimensions of Information Quality in their Effects on Healthcare Quality in Hospitals (Byrd & Byrd, 2013)	Accuracy, Completeness, Timeliness	M
Efficient quality-driven source selection from massive data sources (Lin et al., 2015)	Completeness, consistency, coincidence	M
Measuring the quality of patient data with particular reference to data accuracy (Gibson, 1997)	Accuracy, Completeness, Precision, verifiability, validity, plausibility	M
Open data quality measurement framework: Definition and application to Open Government Data (Vetro et al., 2016)	Accuracy, Completeness, understandabilit	M

	y, Traceability, compliance	
Does use of computer technology for perinatal data collection influence data quality? (Craswell et al., 2016)	Accuracy, consistency, clarity	M
Identifying Relationships of Information Quality Dimensions (Lee & Haider, 2013)	Believability, security, accuracy, timeliness	L
Review of data quality dimensions and applied methods in the evaluation of health information systems (Lima et al., 2009)	reliability, validity, coverage, accuracy, completeness	M
Early Childhood Chronic Illness: Comparability of Maternal Reports and Medical Records (Miller et al., 2001)	Consistency, accuracy, plausibility	M
The Challenges of Data Quality and Data Quality Assessment in the Big Data Era (Cai & Zhu, 2015)	Availability, usability, reliability, relevance, presentation quality	M
The Effects and Interactions of Data Quality and Problem Complexity on Classification (Blake & Mangiameli, 2011)	Accuracy, Completeness, Consistency, Timeliness	L
The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases (Salati et al., 2016)	Completeness, reliability	L
The use of big data in manual physiotherapy (Rhodegero, 2014)	Completeness, accuracy	M
Transparent Reporting of Data Quality in Distributed Data Networks (Kahn et al., 2015)	Format, availability, timeliness, consistency	L
A Pilot Ontology for a Large, Diverse Set of National Health Service Healthcare Quality Indicators (Pam, 2014)	Consistency, conciseness, completeness, expandability, sensitivity.	M

Ontological Multidimensional Data Models and Contextual Data Quality (Bertossi & Milani, 2018)	Consistency, Currency, Accuracy, Completeness, Redundancy	L
Quality Awareness for a Successful Big Data Exploitation (Capiello et al. 2018)	Accuracy, Completeness, Consistency, Distinctness, Precision, Volume	M
An Introduction to Dynamic Data Quality Challenges (Labouseur & Matheus, 2017)	Accessibility, Ease of manipulation, Representation	L
Big Data, Big Data Quality Problem (Becker et al., 2015)	Accuracy, Consistency, Completeness, Timeliness, Pedigree, Precision, Relevance	M
A Model for Addressing Quality Issues in Big Data (Onyeabor & Azman, 2018)	Accuracy, Consistency, Completeness, Timeliness	M
Data Quality in Big Data Processing: Issues, Solutions and Open Problems (Zhang et al., 2017)	Accessibility, Timeliness, Credibility, Accuracy, Consistency, Integrity, Completeness, Fitness	M
Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review (Ji et al., 2020)	Correctness, Completeness, Timeliness, Accuracy, Consistency	M
Towards a Data Quality Framework for Heterogeneous Data (Micic et al., 2017)	Completeness, Consistency, Uniqueness	M
Review of Factors Influencing Patient Readmission Based on Data Quality Dimension Model (Mohmad et al., 2020)	Completeness, Timeliness,	M

	Accuracy, Consistency	
Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses (Spengler et al. 2020)	Conformance, Completeness, Plausibility	M
Big Data and Quality: A Literature Review (Lakshen et al., 2016)	Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency	L
Data quality in ETL process: A preliminary study (Souibgui et al., 2019)	Duplication, Completeness, Accuracy, Interpretability, Conciseness, Consistency	L

Each weight was assigned a numerical factor as follows: L: 1, M: 2 and H: 3. Low weights were assigned whenever the research lacked both Big Data and health industry context yet included discussions about DQDs. Medium weights were assigned when, over and above DQDs discussions, either a Big Data or health industry context was present. High weights were assigned when both concepts (Big Data and health industry) were present. Inevitably, the judgement of the researcher towards the overall alignment of papers compared to the context of the current research also affected the assignment of weights. A weighted total was computed to ascertain which DQDs were the most important, following the application of the inner hermeneutic cycle. A staggering unique count of 52 distinct DQDs was noted following the IHC analysis. This confirmed the impression of a lack of a universal DQD framework and the possible fact that different authors might be using different jargon to express the same idea. To overcome the difference with jargon is some cases, this research study mapped ideas/jargon expressing a DQD to a more conventional term expressing a DQD, and hence assuring the validity of DQDs discussed with the IHC. As the research objective was to investigate the most important DQDs for Big Data in health industry, the DQDs cited only once with a low weight were discarded from further future analysis, and the remaining number of DQDs became 46. Count results of DQDs listed tended to confirm previous general research studies on DQDs with accuracy being one of the most cited.

The results are displayed in Table 4.3 below, sorted in descending order:

Table 4.3: Weighted Counts of different DQ dimensions

DQ dimension	Total count	DQ dimension	Total count
Accuracy	73	Readability	2
Completeness	71	Capture	2
Consistency	47	Privacy	2
Reliability	15	Heterogeneity	2
Timeliness	22	Provenance	2
Currency	10	Comprehensiveness	2
Availability	6	Concordance	2
Accessibility	8	Confidence	2
Trust	5	Data volume	4
Security	5	Coincidence	2
Correctness	7	Verifiability	2
Plausibility	6	Understandable	2
Relevance	6	Traceability	2
Clarity	4	Compliance	2
Validity	5	Coverage	2
Uniqueness	6	Usability	2
Format	3	Presentation quality	2
Precision	7	Expandability	2
Usefulness	3	Sensitivity	2
Distinctness	2	Pedigree	2
Credibility	2	Integrity	2
Fitness	2	Duplication	2
Redundancy	2	Conformance	2

It is clear from the count analysis of Table 4.3 that DQDs such as ‘accuracy’ and ‘completeness’ are the most important within existing studies with 73 and 71 counts. The current research study posits that the DQDs with a total count of more than 10 are the ones which are most important in the context of big data within the health industry. Therefore, the most important DQDs were *accuracy, completeness, consistency, reliability and timeliness*.

Analysing the results with the perspective of the framework of Wang and Strong (1996) would result in Table 4.4 below:

Table 4.4: DQ category dimensions with count aggregate

Category	Individual dimensions	Count
Intrinsic	Accuracy, trust, plausibility, precision, compliance, traceability, verifiability, provenance, confidence, concordance, correctness	112
Contextual	Completeness, timeliness, currency, reliability, availability, uniqueness, relevance, validity, expandability, sensitivity, coverage, data volume, comprehensiveness, heterogeneity	153
Representational	Consistency, format, usefulness, readability, capture, coincidence, understandability, usability, presentation quality	65
Accessibility	Accessibility, security, privacy, compliance	17

Hence, the following may be concluded for each of the research questions:

RQ1: Should intrinsic DQ also be applicable for Big Data in the health industry since everywhere there is use of data, those DQ dimensions are supposed to be needed?

The intrinsic category ranks second as per Table 4.4. This supports the opinion that some DQDs such as 'accuracy' and 'trust' are applicable in all situations where data could be used, including that of Big Data for the health industry. But at the same time, the fact that intrinsic DQ category is not the most cited category gives some credit to research which state that data quality might not be impactful in the context of Big Data. Finally, it also indicates that data quality studies in this current context should not be considered intrinsically but the different applications and user point of views also impact upon the importance of DQDs.

RQ2: Would contextual DQ be highly relevant for Big Data in the health industry, as different DQDs could be applicable based upon the context of the use of data?

The contextual category carries the highest importance. The result of the IHC determines that those DQDs constituting the contextual category have a higher collective importance for Big Data in the health industry. This might be explained by the fact that there is such a huge variety of categories of data (patient-related, genomic, trauma based, and others) and so many different end-consumers of data (doctors, insurance companies, pharmaceutical groups, and others) such that each specific use of data might uphold different perspectives of quality to suit the "fitness for use" definition of data quality. The conclusion is that the context is extremely important for data quality applications in the specific domains of Big Data within the health industry. Hence, future data quality initiatives for this research would need to focus upon specific use cases or datasets.

RQ3: Representational DQ could be less applicable for Big Data in general due to the variety characteristic and the fact that the meaning of data would depend upon the aims of data analysis. Is representational DQ less important for Big Data in the health industry?

The importance of the representational DQ category is quite low relative to the two previous categories with a total count of only 65. The hypothesis was that this category could be relatively unknown due to the variety characteristic of big data. However, most of the research data involved in our IHC concerns mostly text-based data, including numbers, and the results were being analysed and used by well-trained personnel. This could warrant future work concerning data quality with other kinds of data (images, charts and videos) used in the health industry. On the other hand, even if some authors such as Caballero et al. (2014) point towards the final goals of data analysis and how data quality were therefore important, most other research studies as part of the IHC do not give enough indication pertaining to the rationale behind the data analysis. Therefore, there is the inherent assumption in most research studies that data would be used in one or very few use cases; as discussed in the earlier sections of this research study, this assumption might not hold true, especially in the Big Data context.

RQ4: Accessibility DQ should not be among the most important category for Big Data in the health industry as in terms of access, the datasets are very often publicly available?

Finally, the accessibility category ranks unsurprisingly last in our finding, which confirms our hypothesis present in RQ4. Many datasets are publicly available for analysis and research; hence previous research studies were undertaken within a context where accessibility was easy. Thus, the volume aspect could be deduced not to have affected the accessibility to data, but care should be taken to probe the investigation about the relationship between volume and accessibility for Big Data applications. Furthermore, for private and industry-based applications of big data within the health industry, it could be argued that DQ dimensions such as security, privacy and compliance would have a higher impact. Thus, future Big Data quality frameworks or methodological approaches specifically for the health industry should explore this accessibility category deeper.

4.4 Implementation of LSA

The application of the IHC might be criticized as being too tightly associated with the interpretation of the author relative to the weights given to the different documents and hence to the selection of most important DQDs in the research context. Therefore, in order to bring some more objectivity in the integrative review process, term to documents inductive relationship

methods was applied in the form of Latent Semantic Analysis. As described in section 3.4, this inductive relationship of term relative to a whole corpus of documents is different to the fields of topic modelling and relevance ranking, where for the latter other algorithms such as LDA and DeepRank could have been considered. The LSA is a proven technique for this specific purpose (Landauer et al., 1998; Harada et al., 2020) and is adequately supported in terms of available algorithms for implementation.

The application of LSA for this research made use of a corpus of 43 research abstracts which were selected from the same corpus which was used previously for the IHC implementation. It was essential to use the same documents involved with the IHC as the aim of applying the LSA was to interpret the most important DQDs within a well-defined group of documents in an objective way. However, due to issues such as some of the articles not having an abstract section, only 43 abstracts could be used. The decision to use only abstracts for detecting relationships relative to terms within documents was adopted from an authoritative work in the field of LSA application (Deerwester, et al., 1990; Casakin and Singh, 2019). The use of abstracts instead of full texts is not only a well-accepted practice with the application of LSA, but is recommended for the following reasons (Casakin and Singh, 2019):

1. They provide self-containing text that summarizes the whole paper but contain less noise from an LSA perspective. The noise could be in the form of figures and tables which are not accessible to semantic text analysis and therefore using the main body of the paper might miss some critical information for the LSA analysis.
2. In the main text, authors could repeat the same themes to emphasize key messages, and this could adversely affect term frequency calculations of LSA.

The steps taken to implement the LSA are discussed below.

Step 1: A corpus of raw text was assembled. Since the documents containing previous literature in this current research study area consisted of 54 documents in pdf format but only 43 contained an acceptable abstract section, the LSA implementation had to be limited to 43 documents. The first step was to apply an algorithm to convert .pdf documents into text format (.txt). This was undertaken using the "pdfminer" class of python. However, during this process, figures and charts were not converted into text, but this did not have any negative impact on the LSA implementation as no abstract contained figures and charts. Furthermore, characters such as '=, <, >' generated a compiler error with the 'Gensim' package. Those characters were removed from the documents inserted as input to the LSA algorithm without any potential consequence on the LSA results as they did not show any link to DQDs. Also, some pdf articles had been

accessed as image-based articles, and therefore could not be directly converted to text format. Hence, a non-pdf equivalent was searched for using research databases and the text equivalent of the abstract was extracted. The total size of the 43 documents were able to be processed on a laptop with 8 GB of RAM. The algorithm was able to identify the documents using a zero-based indexing system. Table 4.5 (below) provides a list of index numbers mapped to research article titles:

Table 4.5: Mapping of index numbers to research article

Index	Research title
0	A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical research
1	Completeness and accuracy of data transfer of routine maternal health services data in the greater Accra region
2	A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments
3	Valid comparisons and decisions based on clinical registers and population-based cohort studies: assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database
4	Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges
5	Big Data Quality: A Quality Dimensions Evaluation
6	A Hybrid Approach to Quality Evaluation Across Big Data Value Chain
7	From Data Quality To Big Data Quality
8	Challenges in data quality: the influence of data quality assessments on data availability and completeness in a VMMC programme in Zimbabwe
9	Classifying, measuring and improving the quality of data in trauma registries: A review of the literature
10	Creating a General (Family) Practice Epidemiological Database in Ireland - Data Quality Issue Management
11	Data Challenges in Disease Response: The 2014 Ebola Outbreak and Beyond
12	Data Mining Consulting Improve Data Quality
13	Data Quality: A Survey of Data Quality Dimensions
14	Data Quality by Contract – Towards an Architectural View for Data Quality in Health Information Systems
15	Data Quality Problems When Integrating Genomic Information
16	Data representation factors and dimensions from the quality function deployment (QFD) perspective
17	Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research
18	A Methodology to Evaluate Important Dimensions of Information Quality in Systems
19	The influence of calibration method and eye physiology on eyetracking data quality
20	Improving the Data Quality of Drug Databases using Conditional Dependencies and Ontologies
21	Contrasting the Dimensions of Information Quality in their Effects on Healthcare Quality in Hospitals

22	Efficient quality-driven source selection from massive data sources
23	Measuring the quality of patient data with particular reference to data accuracy
24	Open data quality measurement framework: Definition and application to Open Government Data
25	Does use of computer technology for perinatal data collection influence data quality?
26	Identifying Relationships of Information Quality Dimensions
27	The Challenges of Data Quality and Data Quality Assessment in the Big Data Era
28	The Effects and Interactions of Data Quality and Problem Complexity on Classification
29	The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases
30	Transparent Reporting of Data Quality in Distributed Data Networks
31	A Pilot Ontology for a Large, Diverse Set of National Health Service Healthcare Quality Indicators
32	Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work
33	Discovering Dependencies among Data Quality Dimensions: A Validation of Instrument
34	Quality awareness for a Successful Big Data Exploitation
35	Big Data, Big Data Quality Problem
36	A Model for Addressing Quality Issues in Big Data
37	Data Quality in Big Data Processing: Issues, Solutions and Open Problems
38	Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review
39	Towards a Data Quality Framework for Heterogeneous Data
40	A Review of Factors Influencing Patient Readmission Based on Data Quality Dimension Model
41	Big Data and Quality: A Literature Review
42	Data quality in ETL process: A preliminary study

Step 2: The text documents were pre-processed for analysis. Firstly, this involved eliminating unnecessary characters such as page numbers, symbols and white spaces. Secondly, a stop list of words that should be excluded from analysis was devised. The complete stop list used was *"for, a, of, the, and, to, namely, higher, in, on, at, data, their, ours, yours, her, his, and, from, other, are, with, such, but, require, is, care, We, we, They, these, using, over, can, that, towards, within, between, known, be, users, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, &, first, were, was, also, %"*. The LSA algorithm developed being case sensitive, therefore some words had to be inserted in both lower and upper cases. Most of the words are common trivial English words which do not have a major difference in the semantic structure of the documents. All of them would not have been discarded by the LSI method which basically creates an index based upon more than one occurrence inside the whole corpus of 43 documents/abstracts. Thus, their specification as terms to be discarded during the first sequel of processing was essential. This stop list could have included many other words which according to human judgement would not have had any kind of major impact upon

the semantic structure of the documents relative to data quality dimensions. However, because LSA does consider the semantic structure in a much more complex way compared to human judgement, such as not just comparing exact query terms but having some global evaluation of the terms within a corpus of documents, the decision was taken to only specify common English terms and special characters. Subsequently, word stemming was performed using pre-built word stemming algorithm known as 'NLTK' with python programming language to standardize some common English terms.

Step 3: A term document matrix was created. This is a row-column tabular representation of counts of terms per document. The columns normally represented the document, which in this current research study referred to existing research studies in the area of DQDs. The rows represent the terms, which referred to the 46 data quality dimensions. Python 2.7 was used, along with the *Gensim* package. A dictionary was built from the 43 documents which resulted into 857 unique tokens, which refer to terms extracted from the documents. This package contains classes allowing the TF-IDF transformation and the dimensionality reduction for the SVD. After application of the IDF weights, a 2809 matrix of non-sparse elements was created.

Step 4: The dimensionality of the term matrix document was reduced. SVD would be applied upon the term matrix document to relate the importance of terms per document. The SVD was applied in two phases; during the first phase, an 857 by 110 action matrix was constructed and subsequently 'orthonormalized'. During the second phase, a dense SVD was carried out and created a 110 by 43 action matrix. Ten factors were kept by the LSA algorithm which resulted into 66.26% of discarding of the energy spectrum (squared sum of eigenvalues), which results into elimination of around two third of terms in the corpus. An example of the output per document is as follows for document 0: topic #0(1.607): -0.385*"big" + -0.142*"information" + -0.135*"assessment" + -0.135*"dimensions" + -0.108*"paper" + -0.102*"health" + -0.102*"have" + -0.099*"framework" + -0.093*"research" + -0.088*"clinical".

Finally, a cosine similarity function was computed upon the index created from the 43 documents with the use of 10 features. For the search query term 'Completeness', the following result was produced:

[(8, 0.90279394), (3, 0.77473396), (29, 0.76323533), (1, 0.7185556), (22, 0.5817686), (30, 0.54503775), (23, 0.49930415), (40, 0.38905212), (5, 0.3672621), (42, 0.3548857), (20, 0.35378885), (2, 0.29887176), (4, 0.29301947), (11, 0.27644438), (18, 0.26651943), (31,

0.21281636), (19, 0.2119858), (9, 0.2049908), (6, 0.20460585), (36, 0.18418416), (0, 0.18377274), (24, 0.16630718), (25, 0.12100372), (10, 0.118424356), (13, 0.08647175), (26, 0.059444256), (37, 0.054796163), (33, 0.0405797), (17, 0.029205628), (12, 0.021762729), (38, -0.013150033), (15, -0.02922683), (27, -0.031857494), (16, -0.034892436), (41, -0.053192116), (35, -0.06098572), (34, -0.062179472), (7, -0.06682982), (21, -0.13487153), (39, -0.14402921), (28, -0.14697781), (14, -0.26347446), (32, -0.35439724)]

The interpretation of the sample value (8, 0.90279394) could be understood as follows; 8 refers to a specific research study bearing index value 8 and therefore titled "Challenges in data quality: the influence of data quality assessments on data availability and completeness in a VMMC programme in Zimbabwe". 0.90279394 refers to the relative importance of the DQD 'Completeness' within this particular document. Therefore, those set of values plot the relative importance of DQD 'Completeness' across all the 43 documents used; for some of them, the importance is very high while at the other extreme, the importance of the term is negative.

4.4.1 LSA Algorithm created

The pseudocode used for the LSA implementation is as follows:

Documents = text abstracts of 43 research articles
Specify the stop list of words
Retrieve all words one by one from the documents
    If a word appears more than once and not part of stop list, then add it as a token.
Create a dictionary with all individual tokens
Apply TF-IDF upon all the tokens
Apply Lsimodel method upon the term document matrix
Specify the matching/search query term
Apply MatrixSimilarity method to generate an index
Sort the index and display

4.4.2 Analysis and findings of LSA

The results presented in Appendix 3 were obtained after the application of the above-mentioned algorithm for the 46 DQDs detected through the IHC method. The rows represent the index number as shown in Table 4.5 above whereas the columns show distinct DQDs. The values represent the cosine similarity of terms, in this case DQDs, to documents, in this case research study abstracts. Cosine similarity represents the importance of a term per document, without taking into account individual word counts of each document. Out of the 46 dimensions used as query documents, 36 did not return any level of similarity nor difference with the document corpus being examined. Hence, for all documents making up the corpus, the cosine similarity indicated 0. The columns representing those dimensions have thus been removed from the table in Appendix 3.

4.5 Principal findings

The first major difference between the LSA and the IHC results is that ***only 10 out of possible 46*** DQDs showed some similarity with the 43 documents forming the corpus with the LSA method. This could be explained by the difference between the semantic interpretation of some text by a human reader and LSA in the perspective of a corpus of documents. Therefore, this current study comes up to with the conclusion that some DQDs such as timeliness, which had a score of 11 for the IHC and 0 for LSA, just did not have sufficient representation in terms of similarity when considering the whole corpus of abstracts merged together. This is because LSA does not work as a full text query for words but rather considers the importance of terms in a holistic interpretation of the corpus. This holistic interpretation might result in some very surprising kind of results such as: (1) some terms which appear a lot in the documents might not result in a high cosine similarity, (2) some terms which does not appear at all in the documents might show some similarity with those documents, (3) some documents which did contain certain terms display zero cosine similarity for some terms and (4) high cosine similarity for some terms per specific document.

The fact that some DQDs did not reflect any match after the LSA application is closer to possibility number 3 discussed above. Again, taking the example of the **timeliness** DQD, which according to the IHC study involved 14 research articles, was the fourth most important DQD for Big Data in health industry, and for the LSA application, 13 out of the 14 article abstracts were referenced. Still, the cosine similarity matches were 0 for all research abstracts. This proves

that with LSA, it is not individual word counts which would be considered, but the semantic strength and relationship of terms within an overall corpus of documents.

The following charts focus upon the importance of the ten individual DQDs which denoted some degree of importance towards the whole corpus. The x-axis shows the different documents provided by their index number as per table 4.5.

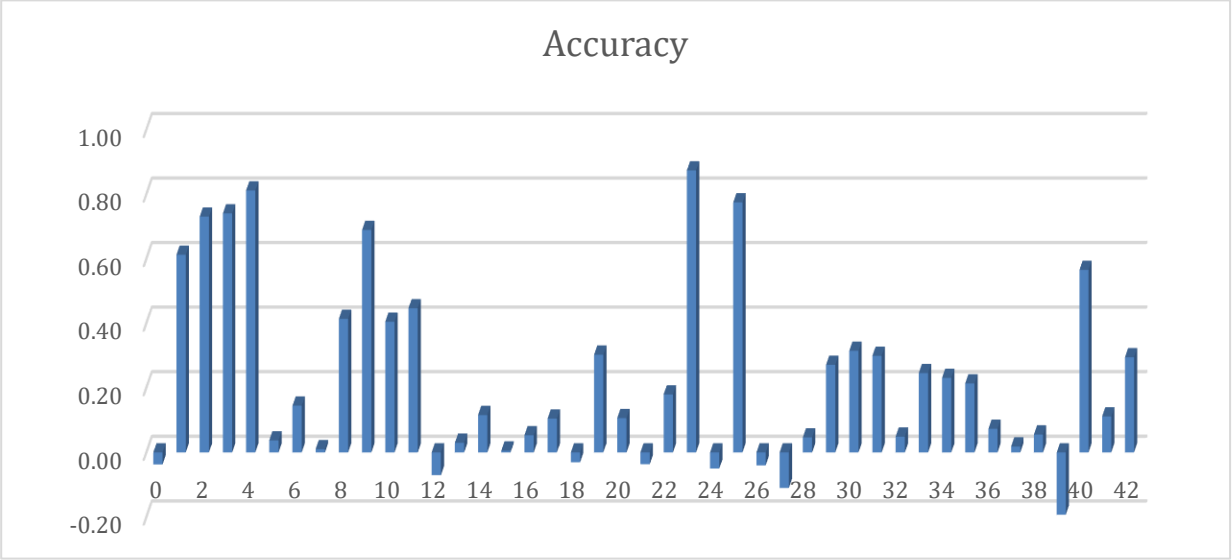

Figure 4.2: Importance of Accuracy DQD per individual article

The accuracy DQD denotes a relatively high degree of importance as in Figure 4.3 above with seven papers having a cosine similarity greater than 0.6.

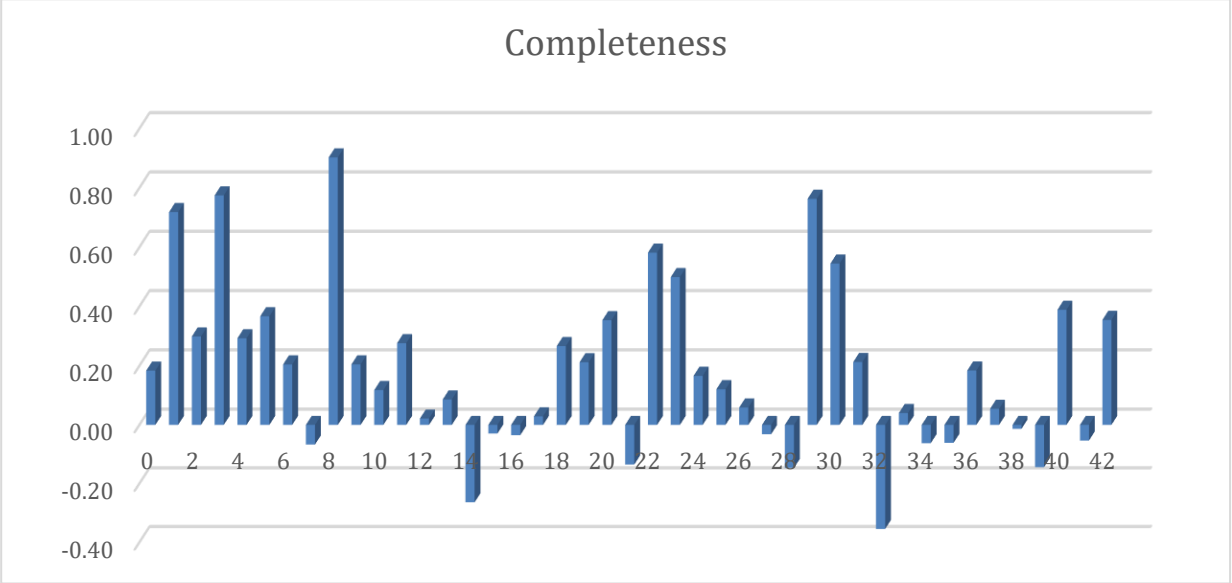

Figure 4.3: Importance of Completeness DQD per individual article

With Figure 4.4, it can be denoted that completeness DQD has only four articles with a cosine similarity greater than 0.6.

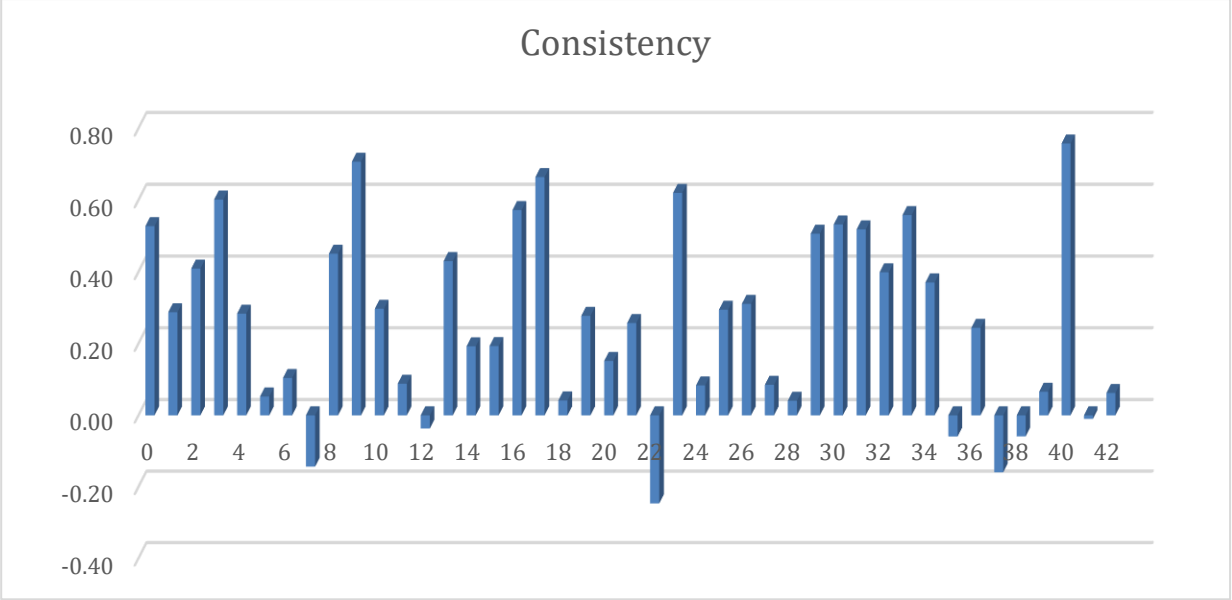

Figure 4.4: Importance of Consistency DQD per individual article

Four articles denote a cosine similarity of greater or equal to 0.6 for the consistency DQD as per Figure 4.5 above.

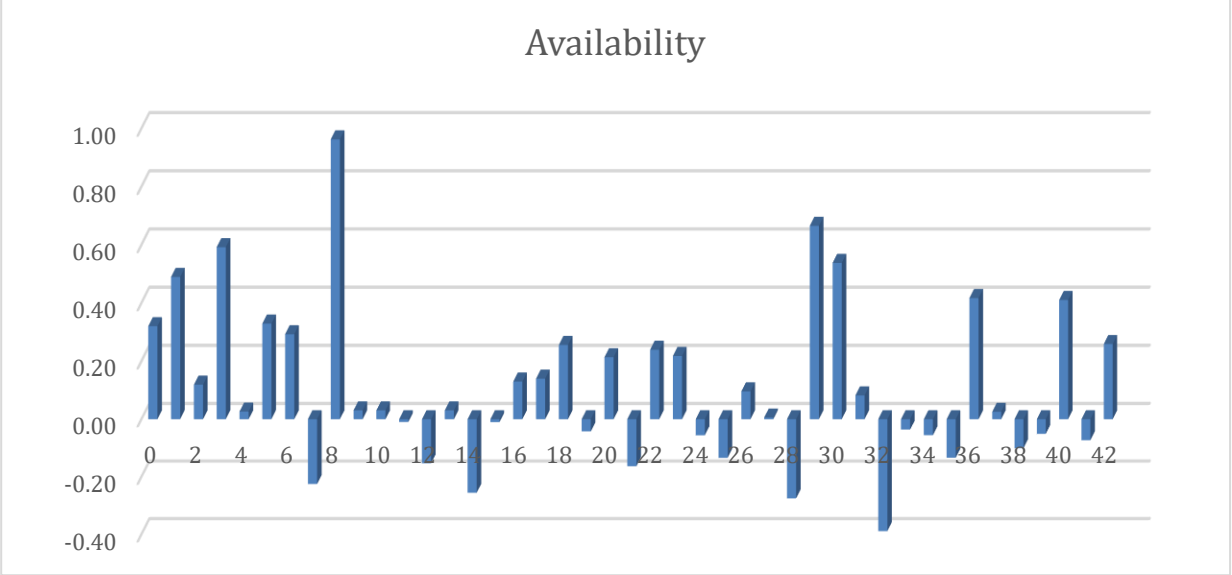

Figure 4.5: Importance of Availability DQD per individual article

The availability DQD denotes only two articles with a cosine similarity of greater or equal to 0.6 and therefore relatively less important compared to accuracy DQD.

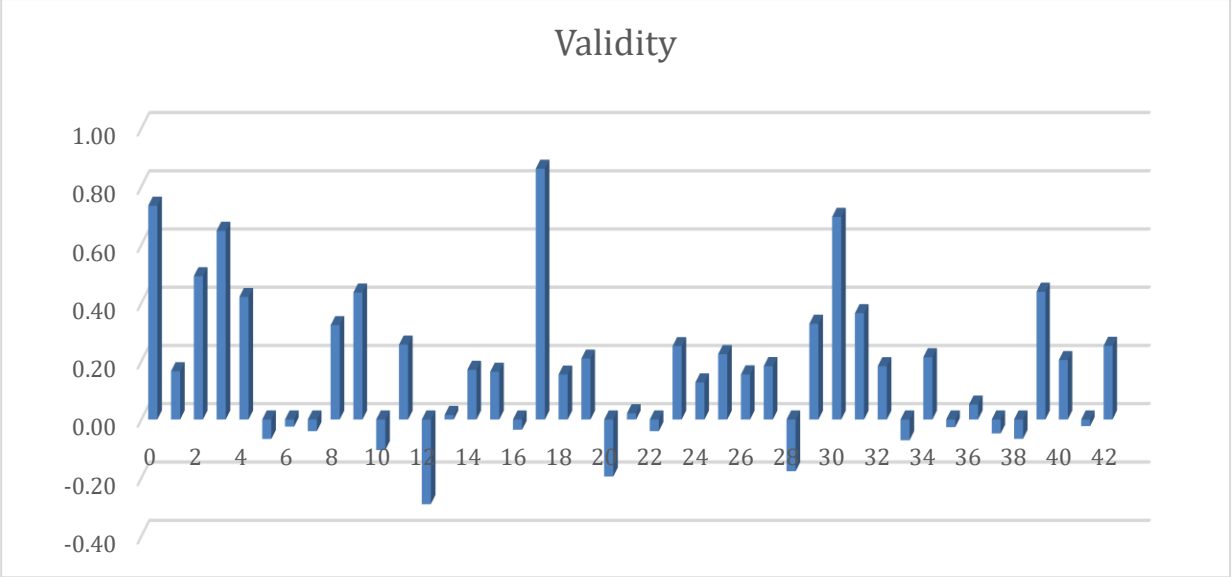

Figure 4.6: Importance of Validity DQD per individual article

Four articles show a cosine similarity of greater or equal to 0.6 for the validity DQD as per figure 4.7 above.

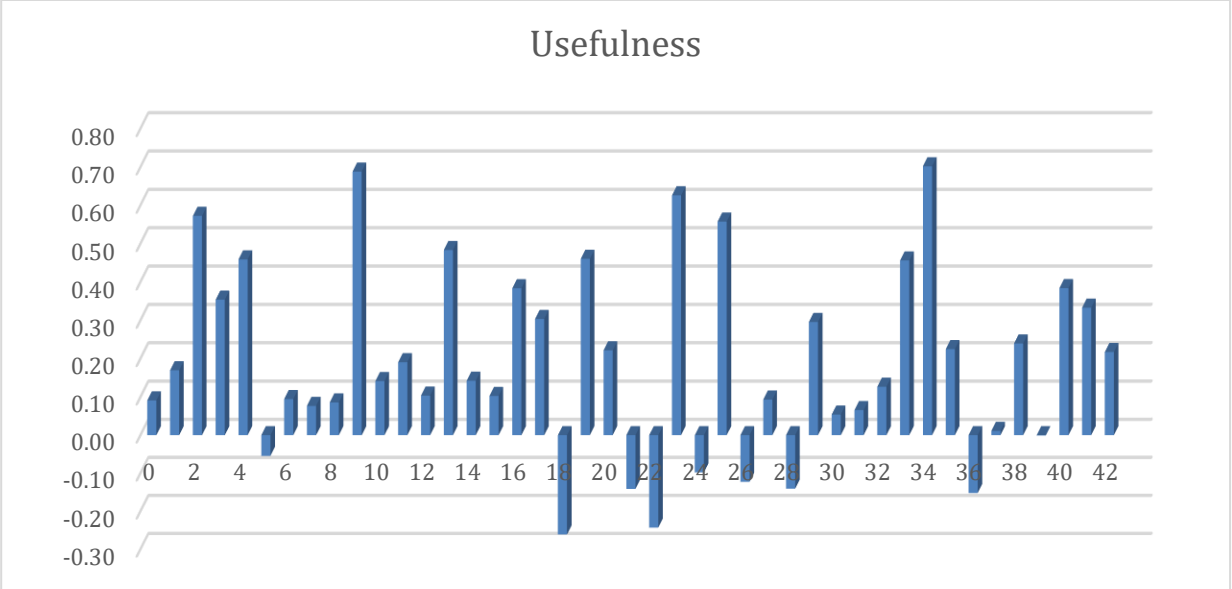

Figure 4.7: Importance of Usefulness DQD per individual article

The usefulness DQD denotes a relatively high importance as five articles show a cosine similarity of greater or equal to 0.6.

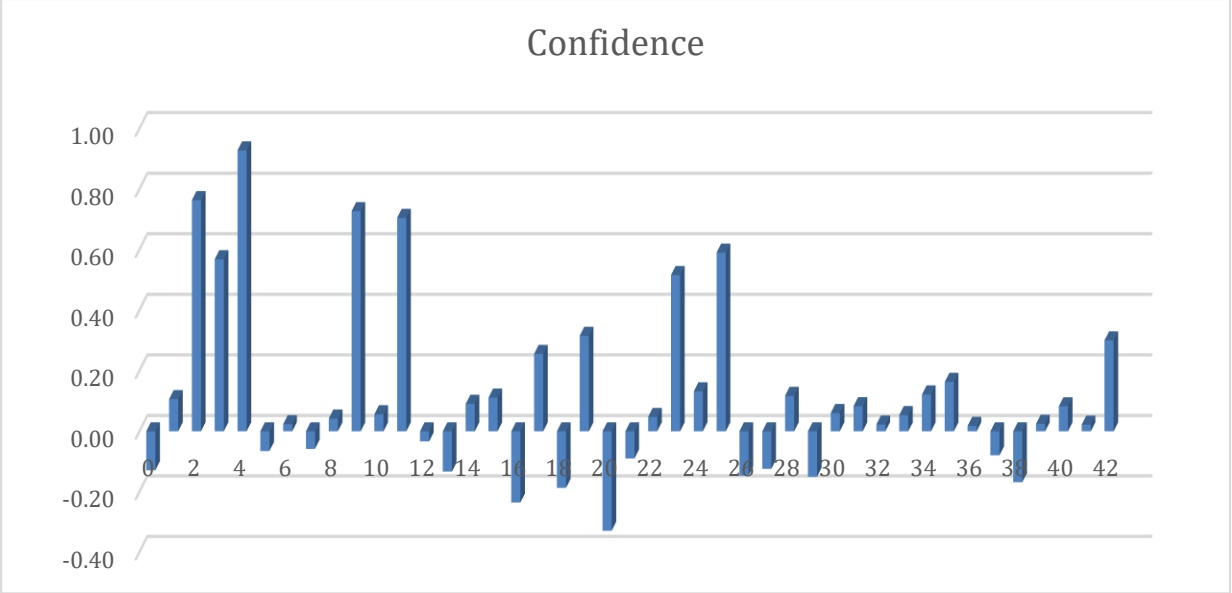

Figure 4.8: Importance of Confidence DQD per individual article

From figure 4.9, it can be seen that the confidence DQD denotes four articles with a cosine similarity of greater of equal to 0.6, and therefore, relatively important.

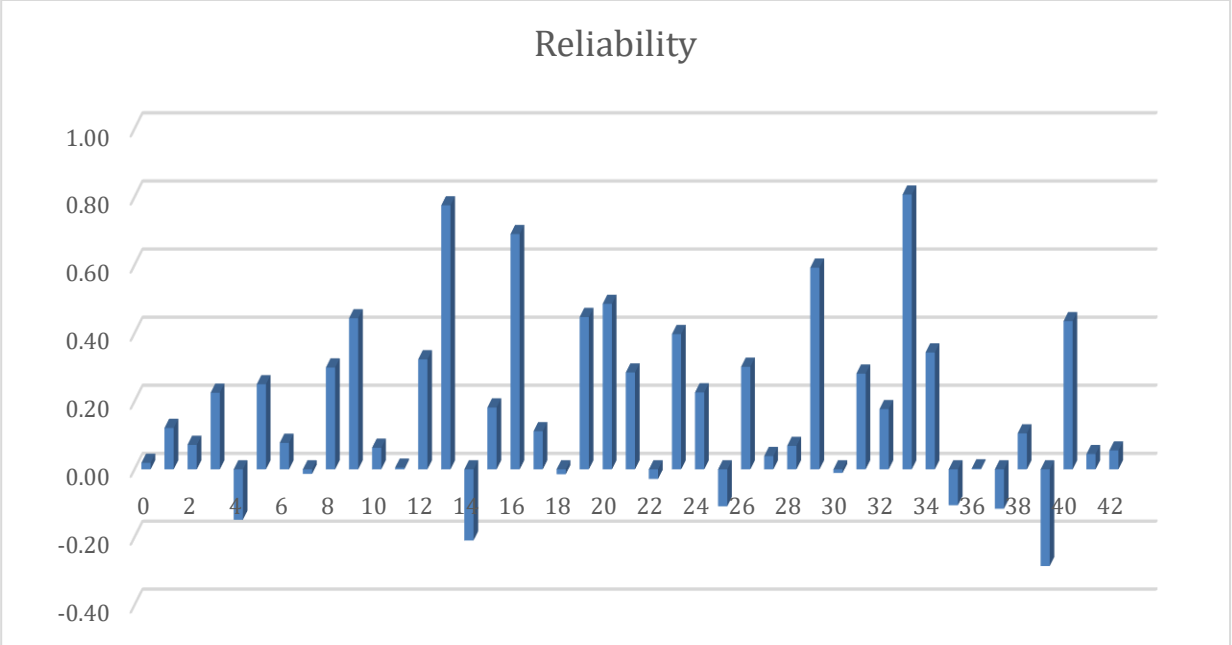

Figure 4.9: Importance of Reliability DQD per individual article

With only three articles showing a cosine similarity of greater or equal to 0.6, the reliability DQD is relatively less important.

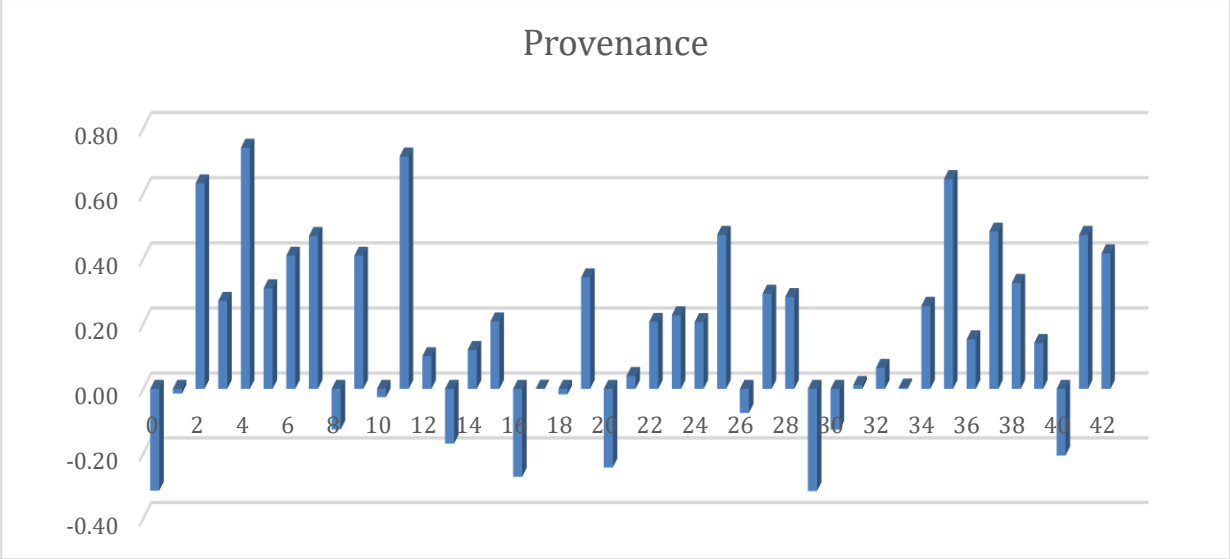

Figure 4.10: Importance of Provenance DQD per individual article

With four articles denoting a cosine similarity of greater or equal to 0.6, the provenance DQD also denotes a relatively fair amount of importance.

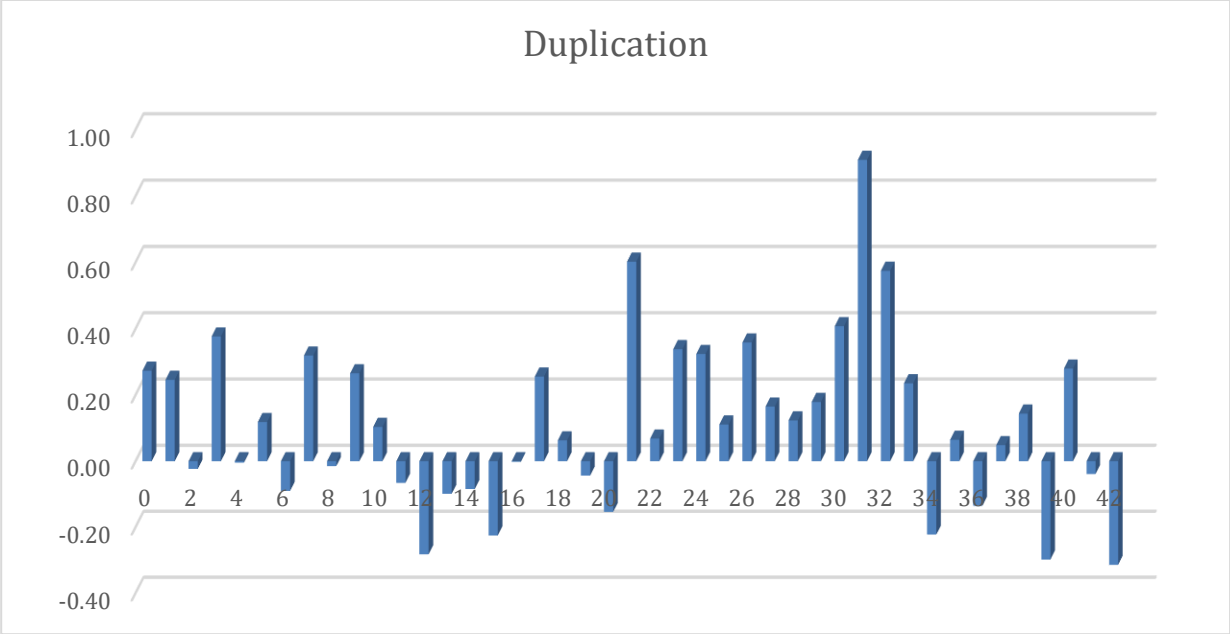

Figure 4.11: Importance of Duplication DQD per individual article

The duplication DQD has only 1 article denoting a cosine similarity of greater than 0.6, and therefore relatively not that important.

The analysis of the individual DQDs per article above shows that the accuracy DQD is the most important but also shows many DQDs with almost same amount of importance. Hence, taking only a subsection of cosine similarity which is greater than 0.8 as benchmark for the most

important DQDs with the cosine similarity index, we denote the following facts in Table 4.6 (below):

Table 4.6: Hierarchy of DQDs per cosine similarity

DDQs	**Counts of cosine similarity >= 0.8**
Accuracy	2
Completeness	1
Consistency	0
Availability	1
Validity	1
Usefulness	0
Confidence	1
Reliability	1
Provenance	0
Duplication	1

Taking the above into account, ***Accuracy*** is again indicated as the most important DQD followed by ***Completeness, Availability, Validity, Confidence and Duplication.*** This hierarchy is confirmed to be different from the IHC method which resulted in *accuracy, completeness, consistency, reliability and timeliness* as most important DQDs. On the other hand, ***accuracy*** is identified as the most important DQD by both LSA and IHC.

Analysing the distribution of the cosine similarity indexes per range bands gives another interesting insight; there is significantly much more similarity plots between the ranges of 0 to 0.4 compared to the range of 0.6 to 1. This means that even some if DQDs have been identified across the corpus of abstracts, their importance is largely ranked from low (0) to medium (0.5). One potential justification of this phenomena could be the fact that most of the 43 research articles discuss data quality, but do not necessarily focus upon DQDs as their main area of research. However, upon comparison with the IHC results, the same trend could be discerned as with the IHC, there were 5 DQDs with a weighted count of greater than 10 and most of the 38 other DQDS had very low weighted counts.

Hence, combining results from both the IHC and the LSA, the two most important DQDs discovered would be undoubtedly *accuracy*, being found most important by both methods and *completeness*, ranked second by IHC and very highly by LSA.

4.6 Conclusion

This research study has been set in a multi-disciplinary context involving three main fields: data quality (as an enabler of data governance), Big Data and machine learning. Although data quality is a well-researched field for the last two decades, there is still a lack of precise nomenclature when it comes to DQDs. Big Data is an emerging field, with few existing research studies focusing on the perspective of data quality. Optimising data quality for Big Data in general did seem too broad in scope, and hence, a specific area of study in the context of health data was initially chosen. Upon further investigation in the field of health data, it was found that the huge number of different types of health data being used still makes the research study broad and not concrete. Subsequently, real word datasets containing data used in a practical context and containing realistic data would be selected for further experiments as per the research objectives of this study.

Literature review indicated a lack of uniformity and standardisation in terms of DQDs in general; DQDs discussions for Big Data in the health industry is very rare and the latter are not supported by solid research methods. Hence, the need to carry out a systematic investigation in this area while applying well accepted research methods. Furthermore, no hierarchy or differentiation in terms of importance of DQDs are discussed in existing research studies. This research mapped the most important DQDs to the model proposed by Wang and Strong (1996) and discussed the importance of the different DQ categories.

A qualitative research approach was undertaken. The inner hermeneutic cycle was adopted as one of the research methods due to the emerging research areas of Big Data and health informatics, but also due to the impracticality of implementing other research methods because of organisational, geographical, legal and ethical constraints. To triangulate the results of the IHC approach and possibly eliminate the human bias, an LSA algorithm was found to be most adequate and implemented on the same corpus of literature as identified by the IHC.

The results of the IHC confirmed the popularity of some well discussed DQ dimensions such as *accuracy, completeness and consistency*. With the IHC, the *reliability and timeliness* DQDs add up to the three afore-mentioned DQDs to make up a list of the five most important ones for the context of Big Data within the health industry. When mapping the results of the IHC with the DQ category framework of Wang and Strong (1996), the contextual DQ category was considered

to be most important. This could be easily explained by the broadness of the three domains involved, where there could be thousands of unique applications of Big Data for the health industry. Thus, for each application, the probability of selecting different DQDs increases.

The results between a statistical process for LSA and human interpretation using IHC denoted marked differences; whereas with IHC, 46 different DQDs with varying importance were found, with LSA only 10 DQDs showed some connection with the overall corpus, but again with varying importance. With IHC, *accuracy, completeness, consistency, reliability and timeliness* were found to be the most important in descending order. However, with LSA, *accuracy, usefulness, completeness, availability, validity, provenance and duplication were* found to be most important in descending order. The two common DQDs found most important by the two different methods are ***accuracy and completeness.*** Thus, the first principal result using combined research methods of IHC and LSA is that accuracy and completeness are the most important DQDS to consider for the first research objective of this study. The two methods also confirmed the fact that most of the other DQDs are not being identified as very important in this particular data and industry context. This is a major difference from the results of Caballero et al (2014) who argued that *consistency* is the main DQD. Compared with the results of Weiskopf and Chunhua (2013) in the field of EHR, completeness is found to be a common important DQD but Weiskopf and Chunhua (2013) did not categorize *accuracy* as an important DQD. However, comparison with the work of Weiskopf and Chunhua (2013) should be undertaken with the knowledge that it is unclear whether their work is in the context of Big Data. The results also contrast with the deductions which could had been made from the research of Raghupathi and Raghupathi (2014) where completeness was not identified as a noticeable DQD, even if accuracy was mentioned.

Hence, the work accomplished lays the foundation for further work in the context of the doctoral research study and has laid the premise of the initial steps as per a proposed methodological approach. Accuracy and completeness have been determined as the two most important DQDs with a strong and systematic research method for the very precise context of Big Data in the health industry. Big Data quality initiatives are considered to be meaningless if they are not framed according to specific DQDs (Firmani et al., 2016). Accuracy and completeness DQDs will serve as a basis to help determine which machine learning algorithms could be more efficient in detecting dirty data, as from now on dirty data would be mostly equivalent to inaccurate and incomplete data. Following this, a data repair algorithm based upon accuracy and completeness

related issues will be further investigated as per the research objectives and to inform the proposed data quality methodological approach.

Chapter 5: Evaluating Suitability of Machine Learning Algorithms to Detect Data Quality Issues for EHR Big Data.

5.1 Introduction

Use of Big Data may be ineffective if the raw dataset used contains data of poor quality (i.e., inaccurate, incomplete, inconsistent and unreliable). Hence the need to perform data pre-processing, standardization and cleansing activities to improve the quality of data. Data pre-processing is considered to consume an excessively high amount of data scientists' time (Garcia et al., 2016). Existing research studies described the use of data mining and statistical based methods to support data pre-processing activities, for example, improving *accuracy* of data by trying to predict and fill missing values in datasets (Ma. et al., 2007). This research study hypothesized that the support of machine learning (ML) might make data pre-processing more efficient with the challenges which Big Data brought upon DQ activities. The use of ML for supporting data quality pre-processing upon Big Data is a novel research area, with limited amount of previous research into this specific area. For this present study, the focus of investigation of the potential suitability of ML algorithms for data pre-processing activities is on two specific types of DQDs, namely *data completeness* and *data accuracy*, as what had been discovered as the most important DQDs in the precedent chapter. Thus, this chapter is catering for the third objective in section 1.4.

As set in section 1.4(c) of this thesis, the emphasis was upon investigating the potential of ML for identifying or detecting some DQ issues related to well defined DQDs. For the data 'completeness' DQD, detection of missing values for an attribute was found to be achievable without the need of ML. Conversely, ML algorithms can be applied to solve 'completeness' issues through data imputation techniques (Yu et al., 2017; Ahmed et al., 2016). Data imputation can be based upon statistical measurements such as mean and mode, where a missing value is replaced by a mean or mode value of a data distribution. However, statistically based data imputation techniques might introduce further errors as these mean or mode replacement values might not be accurate and therefore might reduce the value for data analytics. Other data imputation techniques aim to predict much more accurate missing values based upon existing data as part of a dataset. These predictions can be supported with the use of ML algorithms.

As for the ‘accuracy’ DQD, several ML algorithms have been proposed to deal with detection of different accuracy issues such as noise and outlier detection (Hariharakrishnan, et al., 2017; Yinghao et al., 2013). Two general approaches have been proposed to deal with the problem of noisy data (Wu & Zhu, 2008): (1) applying data cleansing methods to eliminate data quality issues as far as possible, and (2) make data mining applications more robust so that they can tolerate the presence of noisy data. The first method presents some drawbacks, such as:

(a) Data cleansing algorithms deal with only certain types of errors,

(b) Data cleansing cannot result into perfect data,

(c) Data cleansing cannot always be applied to all data sources,

(d) Eliminating noisy data may lead to crucial data loss for further mining/analytics

(e) The data mining/analytics algorithm cannot consider the original data source context after data cleansing has been applied.

The points listed above represent limits of data improvement stage for a data quality methodological approach when following a data driven strategy, as this current research study is doing. However, the second approach, that is making data mining applications more tolerant towards the presence of noisy data, is based upon the assumption that there is enough knowledge of the nature of noise that are present as part of a data source. This might hold true in several cases (known device errors, known information transformation errors), but lack of knowledge of data errors due to external data sources used is a very high possibility with Big Data. Hence, the second approach as stated above by Wu and Zhu (2008) holds much more limits in the context of Big Data.

The contribution to knowledge of this section is to compare and evaluate which existing ML algorithms is more appropriate for dealing with missing data and detecting outlier data in the context of EHR Big Data.. A prior assumption discussed in Chapter 3 (Methodology) stated that unsupervised learning category of ML algorithms might be more appropriate for data quality support in the context of Big Data. This chapter will give a more in-depth discussion of this assumption. Ensemble learning techniques were not considered as they are based on a combination of weak learners; the knowledge from this research might in fact propose which weak learners should form part of a new ensemble learning model. The method used for this chapter was based on 1) selection of most adequate existing algorithms from related past studies and 2) experiments involving implementation and execution of identified ML algorithms upon real-world EHR datasets over a cloud infrastructure.

This chapter first discusses the current state of knowledge pertaining to the use of ML connected with data completeness and accuracy issues in general, as there is a limited amount of research for the precise field of detection of data accuracy and data completeness for EHR Big Data. Next, this state of knowledge is validated by applying relevant ML algorithms (discussed within the literature below) on three EHR datasets representing Big Data. In this process, both failed and successful experiments are expanded, as both are equally useful in the formation of a proposed DQ methodological approach. Finally, clear and accepted evaluation measures and metrics are applied to identify the most suitable ML algorithms. The knowledge derived from this chapter was used to inform a proposed data quality methodological approach for the overall thesis.

5.2 Literature review

5.2.1 Completeness DQD

The *completeness* DQD is expressed in different ways, such as missing values, absent values and sparseness of values (Ahmed et al., 2016). One of the main goals of data pre-processing is catering for missing values to provide high quality data and ultimately maximize value from any subsequent analytics process. The use of the **Bayesian isotonic regression algorithm** was proposed in a past research study for medical data cleaning focusing on blood tests data (Ahmed et al., 2016). The hypothesis was that missing values can be filled with predicted values based on historical data. From that research study, some aspects remain unclear such as the correctness of the imputed values used to replace the missing values, and the percentage of missing values being imputed with this algorithm. One of the generally reported issues of data quality repair techniques is the possibility of new errors being introduced while correcting detected errors (Rahman et al., 2012). Therefore, it is important to have evaluation mechanisms to report on the quality of the imputation process for missing values, and on data repairs in general. As it is one of the very few existing research studies involving imputation of missing values in the health domain, Bayesian isotonic regression algorithm was considered as part of the experiments to evaluate its suitability in the context of EHR Big Data.

Most services offered by Intelligent Transportation Systems (ITS) depend on accurate and complete data (Chen et al., 2018). For ITS, missing values originate mostly from issues with sensing equipment and transmission network, with up-to 56% of reported missing data in an ITS used in Melbourne (Australia). Different missing value imputations (MV) methods have been

proposed throughout existing research studies citing principally ML algorithms such as K-Nearest Neighbours (KNN), Singularity Value Decomposition (SVD), Probabilistic Principal Component Analysis (PPCA) and Low Rank Matrix Completion (LRMC). Chen et al. (2016) proposed a novel self-representation based matrix completion approach for missing data recovery by incorporating l_p-norm regularised sparse self-representation (SRS-l_p). The value of p ranging between 0 to 1 makes the problem non-convex, which is represented as regularisation. Traffic flow data is spatially and temporally correlated with one another, which lends itself favourably towards solutions such as a self-representation based matrix. ***However, these correlations might not exist for the health data context as part of this current research study.*** In Chen et al.'s study, many ML algorithms discussed above, such as KNN, PPCA and LRMC were compared with SRS-l_p. Missing values were artificially introduced in different conditions as there was precise knowledge of the datasets and a root mean squared error (RMSE) was used to measure missing value recovery performance (Chen et al., 2016). The results showed that the Local Least Squares (LLS) algorithm was more effective with low missing ratios, but with higher missing ratios, the self-representation matrix method produce better results. Hence, as part of experiments for this research objective, SRS-lp was considered.

Imputation of missing numerical values are reported to be undertaken efficiently by statistical processes throughout many existing research studies. In a past research study involving computer network-based data, different statistical methods to impute missing values were proposed (Loh and Dasu,2012), such as:

- Random imputation from last 30 measurements
- Random imputation from last 30 measurements of second differences
- Impute using previous non-missing values
- Impute average of last 3 non-missing values

These methods are suitable in datasets consisting primarily of numerical missing values, which could work well in some settings of health data such as laboratory data but will not be sufficient in cases of EHR data containing text data or mixture of different types of data. Furthermore, the precision of imputation using the above mentioned can be questioned. Thus, despite very frequently cited, this research study did not evaluate statistical techniques as part of the experiments for imputation, due to the above reasons cited in this paragraph and also the fact that

the main aim for this objective is focused on the support that ML algorithms might offer for DQ activities.

Yu et al (2017) proposed a modification of a K-NN imputation algorithm, known as Cluster-Based Best Match Scanning (CBMS), in terms of improved computational complexity and improved space/memory usage with comparable level of accuracy to K-NN. The principal logic is that the regression to find potential replacement for missing values is performed not on the whole dataset, but upon clusters of the dataset. As K-NN regression is computationally very expensive with Big Data, it becomes more achievable when the data is broken down into smaller clusters. The key parameter to consider is the number of clusters, and according to Yu et al (2017), it should be large enough but not too large such that the size of each cluster becomes too small. Simulation was carried upon a large smart meter reading dataset and imputation testing accuracy was measured using the mean absolute deviation method. Over and above the computational complexity and memory usage improvements, CBMS proved to be an algorithm which can work with parallel computing. Hence, CBMS was considered for the experiments for this research objective as it was based on a large dataset, even if not related to health data.

There is a multitude of different ML algorithms discussed throughout existing research studies but have not been considered for experiments as some details are lacking about them or they clearly demonstrate certain limitations. However, knowledge about them might inform overall knowledge in this domain. An example of such algorithms is random hot deck imputations, coupled with oversampling and bootstrap methods that was discussed as being more effective in clinical datasets with different types of missing values deficiencies (Lin et al., 2014). However, the experiments carried out made use of small datasets on hundreds of tuples only. Hence, it is quite unclear what would be the performance of random hot deck imputations for Big Data scenarios. The clustering based random imputation (CRI) method was proposed to overcome lack of efficiency of other imputation techniques such as Nearest Neighbour imputation (Zhang, et al., 2006). The original dataset used by Zhang et al. (2006) was divided into two: one without missing values and one containing all instances of missing values. Those datasets were further subdivided into clusters using K-Means to group instances into clusters. Each instance with missing values was matched to comparable clusters containing complete values using Euclidean distance calculation. A kernel-based function was used to impute the missing values in one instance using the comparable information coming from instances and attributes of the matched cluster without missing values. The studies of the techniques discussed above lacked sufficient

details of their implementations and evaluations, and therefore warrant further investigations but could not be implemented as part of experiments for this research objective.

The use of clustering techniques combined with feature selection was promoted to reduce the computation time of both single and multiple imputation methods, while also improving accuracy for data classification tasks (Tran, 2018). The differences between this current research study and the one undertaken by (Tran, 2018) are: (1) the goals of the latter was to improve completeness DQD for further data classification activities, whereas this current research study aims at improving the level of incompleteness of a dataset, irrespective of its subsequent use, (2) the datasets used by (Tran, 2018) were very small datasets, whereas this current research study situates data quality in a Big Data context, and aims to carry out experiments upon datasets demonstrating the volume characteristic of Big Data. However, despite the differences denoted, the reduction of computation time was a very worthwhile feature in a Big Data context, and therefore, the use of clustering machine learning algorithms combined with feature selection was included as part of experiments.

Deep learning techniques, derived from neural networks, are considered for the imputation of missing values in recent research studies. Generative Adversarial Imputation Nets (GAIN) is based on generative adversarial networks (GANs) for imputing missing data. GANs are used to a large extent to work with image-based data and typically makes use of two models, known as the generator and the discriminator (Kim et al., 2020). With GAIN, the generator seeks to accurately impute missing data whereas the discriminator's goal is to distinguish between observed and imputed components. In the process, the discriminator minimizes classification loss and the generator maximizes the discriminator's misclassification rate, resulting in an adversarial interaction (Yoon et al., 2018). GAIN has also been experimented to impute missing data in real-world clinical datasets with mixed variable types (Dong et al.,2021). The comparison of GAIN with two other algorithms, namely 'missForest' and 'MICE' revealed that GAIN was more accurate for imputing data with various level of simulated missingness rates but also that GAIN had a much higher computational speed. However, the details showed that GAIN took an average of 32 minutes to impute approximatively 50000 missing data values. GANs are known to use a high time complexity due to learning and using two models but the experiments by Yoon et al. (2018) pointed towards the improvements in the accuracy of imputation by GAIN as both the amount of data and number of feature dimensions increased. Hence, as part of this current

research study, it was important to include GAIN as part of the experiments for imputation of missing data.

5.2.2 Accuracy DQD

The accuracy DQD can be represented by some specific DQ errors at the instance level of a database. Examples of these errors are missing data, incorrect data, misspellings, ambiguous data, outdated temporal data, "misfielded" values and incorrect references (Laranjeiro et al., 2012). Note the dual categorisation of 'missing data' both as a completeness and an accuracy error, which is a frequent example of lack of standard jargon in the data quality domain. Due to the variety of 'accuracy' errors, there is a need to investigate whether a single ML algorithm could efficiently detect all the types of errors mentioned above. Some of these DQ errors are semantic in nature, such as "misfielded" data, which refer to data values that are inserted in improper data attributes. An example of this would be a first name value inserted in a Surname data attribute. Thus, these semantic errors are difficult to detect automatically and might require some level of human intervention to validate proposed errors detected by automated systems/techniques.

Probing deeper into the accuracy DQD reveals that there are different types of accuracy (Laranjeiro et al., 2012). *Structural accuracy* refers to the general idea of the closeness of a value to a real-life phenomenon. *Syntactic accuracy,* which is a sub-type of structural accuracy, refers to the closeness of a data value to elements in a corresponding domain. For example, even if the true author of a book might be Mr John, but it is recorded as Mr Jack, it might not be syntactically inaccurate if the value 'Jack' forms part of acceptable domain of names. Finally, *semantic accuracy,* another sub-type of structural accuracy, refers to the closeness of a data value to its true value. An example of this would be whether Mr John is the real author of book 'X', and thus refers mostly to incorrect data, which matches closely the correctness DQD (Firmani et al., 2016). This shows the differences of terms used to denote DQ issues within existing research studies. The rapidity with which changes in real-life phenomenon is updated upon data values is known as *temporal accuracy.* Hence, detection of those different types of accuracy issues might be assumed to need different techniques. This current research also investigates whether a single ML algorithm could efficiently address different types of accuracy errors or whether different ML algorithms are required, hence making accuracy issues detection a computationally expensive task.

Some authors (e.g., Rahman et al., 2012) criticised the use of the 'class attributes' technique as the foundation method for determining noisy values either as part of attributes or records in a dataset because there are always some exceptional records that cannot be classified correctly. The same authors further assert that there is typically a low amount of noise as part of datasets. This statement contradicts ideas proposed by most authors in the data science community who argue that datasets typically contain a high amount of errors and therefore require much effort by data scientists for data pre-processing activities. Rahman et al. (2012) proposed an ML based technique known 'CAIRAD: A Co-appearance based Analysis for Incorrect Records and Attribute-values Detection' which depends upon correlation of noise between attribute values; however, Rahman et al. (2012) further explained that many errors, such as typo errors, are random and independent and therefore no correlation of noise exist. Finally, 'CAIRAD' is intended to work only on numerical attributes. Hence, due to all the reasons cited above, there is a need to investigate deeper how far ML algorithms could detect data accuracy issues where there is no correlation between the data points.

A comparative analysis of three machine learning algorithms namely Support Vector Machine (SVM), naïve bayes and Gradient Boosting Decision Tree (GBDT), was undertaken to detect data faults in wireless sensor networks (Yuan et al., 2018). There exist many reasons why sensor nodes might produce faulty data, and those include harsh environmental conditions and poor calibration of sensors, among others. The faults were subdivided into three sub-types, namely noise, fixed and short-term faults; these sub-types were artificially introduced into the experimental dataset used for that research. As part of the evaluation of the three ML algorithms, true positive rate (TPR), false positive rate (FPR), detection accuracy (DA) and precision as benchmarks were used. GBDT outperformed the two other ML algorithms for the three sub-types of faulty data under consideration. Yuan et al. (2018) did not use Big Data, however, as GBDT was explained to outperform SVM and naïve bayes, it was planned to be included as part of experiments.

The presence of 'noise' as part of datasets may result in DQ accuracy problems. The main methods used to detect and remove noise in non Big Data contexts are binning, clustering and regression (Hariharakrishnan et al., 2017). *Binning* involves smoothing out values in a group (bin) by substituting some 'noisy' values with the mean or median value of the bin. The validity of using the smoothing value becomes highly questionable and difficult to implement in the case of real time streaming data, as could be the case for Big Data; without a fixed number of values,

the mean/median value keeps on changing. *Clustering* involves detecting irregular pattern in a dataset. The issue is that this technique is efficient for datasets containing homogenous data, which might not be the case for Big Data due to the variety characteristic. With *regression*, noise in the data is smoothed out via the use of a proper smoothing algorithm (linear, multiple or logistic). A general point of concern is the cost of data cleaning for Big Data, and therefore prohibitive for complex classifiers (Hariharakrishnan et al., 2017). Thus, apart from the accuracy and recall evaluation measures, processing time was considered as another evaluation criteria for this current research study.

Typos represent another category of data accuracy issue. Neural networks classifiers, coupled with knowledge bases, have been proposed as an efficient method to detect typos (Yinghao et al., 2013). The knowledge bases involved in the research (Yinghao et al., 2013) are general English dictionaries such as 'WinEdt', commonly misspelled words aggregated together by Wikipedia and domain specific lexicon regarding vehicle diagnostics. *This introduces the idea that eliminating typo errors might require the use of a reference dataset of correct health terms.* The neural network is trained with a set of misspelled words and their correction candidates. This step is useful to select the most precise replacement whenever a typo had been detected. Experimental evaluation against 'Google Spell' checkers and 'Aspell Check' shows a much better performance in terms of rate of detection and more precise corrections by the proposed system (Yinghao et al., 2013). The current research determined that the methodology used by Yinghao et al. (2013) is not realistic for the Big Data context, as the existence of a dictionary or external knowledge base compatible with the terms in a health data source is not guaranteed to be available.

Sporleder et al. (2006) proposed vertical and horizontal error correction methods as part of semi-automatic error detection tools for text data. Horizontal error correction aims at identifying and correcting errors within a database record whereas vertical error correction aims at doing the same for values inserted in incorrect columns, which was earlier described as 'misfielded' errors. The methods used were data driven and language independent, which expands their range of applications. Also, even if supervised machine language algorithms were used, the authors claimed that there is no need for additional manual annotation since the training set would be obtained from the database itself. *This fact is highly interesting and relevant for this current research study, which entails that it might be appropriate to use supervised learning algorithms even without further manual data annotations.* Precision and Recall were used as evaluation

measures and the results provided very satisfactory results. The test database used as part of the evaluation was quite voluminous and highly dimensional, which again is very similar to what this current research is also aiming for. The techniques used are association rules for horizontal error detection and TF-IDF for vertical error detections.

5.2.3 Summarization of existing knowledge

The following summarizes the most salient points coming out from the literature review. It is to be noted that most of the discussions concerning completeness issues refer to imputation of missing data based on ML methods. Thus, the conventional agreement in this domain is that detection of completeness DQ issues is a straight-forward statistical measurement, and it is the data reparation process in the form of imputation which might benefit from the use of ML.

Table 5.1 lists the ML algorithms selected from the literature review, their key characteristics for the current research study and how far implementation details are available from the existing studies consulted to allow reproducibility in different research contexts.

Table 5.1: Characteristics of ML algorithms from literature review

ML Algorithm	Key characteristics for current research	Support for reproducing algorithms
Bayesian isotonic regression	Has been applied in medical data cleaning contexts; imputation done based on available historical data; no details known about efficiency of imputation of this algorithm	Pseudocodes given, but not enough details upon application of algorithm used.
lp-norm regularisation (SRSp)	Applied upon spatially and temporally correlated data; evaluation demonstrated that SRSp achieve higher missing value imputations in datasets with high amount of incompleteness compared to other ML methods such as KNN. PCCA and LRMC.	Pseudocodes available, together with some mathematical functions. However, lack of precisions upon some variables and functions make it difficult to replicate.
Cluster-Based Best Match Scanning (CBMS)	Modification of K-NN imputation, where K-NN is applied only on clusters and not on all observations; pair-wise, instead of the classical 'pearson' correlations, are used for clustering improved computational complexity and improved space/memory usage	Pseudocodes given but does not seem complete; heavily based upon K-NN associated with clustering.
Clustering combined with feature selection and imputation	General aim is to reduce computation time. Applied both deterministic imputation (DI) and random imputation (SI) methods	Pseudocodes available; used C.50 algorithm based upon different amounts of clusters, and applied DI and SI

Generative adversarial Imputation Nets (GAIN)	Relies upon the adversarial concept to learn the best imputation method through the Generator and discriminator models of GANs. A further mask matrix termed as 'Hint' matrix support the learning process.	Source code available. Unclear of the computational complexity of GAIN for Big Data.
Gradient Boosting Decision Tree(GBDT)	Was used to detect noise in wireless sensor networks; Not applied in Big Data context	Described as iterative decision tree algorithm based on CART regression tree. No other implementation details provided.
Clustering algorithms	Works well with homogeneous data	No details provided; generic implementation considered
Regression algorithms	Outliers smoothed out via adequate smoothing algorithm (linear, multiple or logistic)	No details provided; generic implementation considered
Association rules and TF-IDF	Used to deal with horizontal (record based) and vertical (attribute based) errors respectively.	No details given; generic implementation considered

The literature review seems to suggest that there are different "most efficient" ML algorithms for data quality issues in different contexts. However, the most frequently discussed category of algorithms are *clustering* and *regression algorithms*. The very recent studies involving neural networks associated algorithms such as GAIN with promising initial results also pushes towards exploring deep learning to support DQ for Big Data. The complexity is that as part of the future methodological approach proposed for the current research study, algorithms which could cater for **both** the accuracy and completeness DQDs would be preferable. The exact implementations of those two ML families of algorithms depend on their adequacy relative to the Big Data context and implementation method chosen. The hypothesis for this chapter relative to the possible superiority of unsupervised learning category of ML algorithms is strongly contested through this specific survey of existing studies.

5.3 Experiment Design

There are several tools available for the implementation of ML algorithms. Some examples are WEKA, RapidMiner Studio and Python libraries such as 'scikit-learn'. After a review of different possibilities, RapidMiner Studio was initially selected since it allows extremely fast and easy ML deployment upon different types of data sources. Furthermore, based on reviews from institutions such as Gartner (2020), RapidMiner Studio is cited as one of the best industry tools for data science and ML solutions.

5.3.1 Datasets considered

Three CSV formatted datasets and 1 Google BigQuery public dataset were selected for carrying out the data quality detection and transformation experiments as they embodied Big Data characteristics and metadata analysis showed some DQ issues. Datasets used for experiments with benchmarking purposes had been categorised as 'real world data', 'simulated data' and 'toy data' (Olson et al., 2017). The first category is derived from a real-world problem; 'simulated data' is concerned with artificially generated data made to look like real world data whereas toy data is also artificially generated but without any emphasis of simulating real-world data. The chosen CSV datasets were freely obtained and downloaded from 'www.healthit.gov' and 'www.healthdata.gov' and therefore can be categorized as real-world data. Due to the inherent difficulty of simulating Big Data, this research was carried out on real world data as far as possible. Furthermore, it was vital for this research study to use datasets which were in their original format and had not been curated or pre-processed beforehand. Hence, the correlations between different data values could be learnt by ML models in order to most adequately impute missing data and detect incorrect outliers.

The title of the first CSV dataset was 'EHR Products Used for Meaningful Use Attestation' and it contained data about vendors, products, US health provider specific data and other general public or non-private data. An online document provided metadata about the dataset and specified the different attributes and the attribute descriptions. Those details were essential to understand data quality issues (and in the context of this research study, accuracy and completeness issues) for all the 23 attributes as part of the dataset. The dataset consisted of 1,048,576 rows of data, combined with the 23 attributes, which can be considered as a large dataset in terms of the 'volume' characteristic of Big Data. The second dataset considered detailed payment relative to medical facilities across all the states of the US. There was rudimentary documentation about the different attributes as part of the dataset, but their names were quite explicit about their purposes. It was made up of 18825 rows of data across 23 attributes. The title of the third dataset was 'Behavioral Risk Factor Surveillance System (BRFSS) Prevalence Data (2011 to present)'. There was available metadata information available to help understand purpose of attributes. The number of rows involved was again 1048576, which was then found to be the maximum amount of data in CSV based file which could be worked with in the desktop computer involved in this research. The number of attributes involved was 26. The BigQuery public dataset is named 'covid19_open_data'. It is made up of 701 data

attributes or columns, contains 16,323,138 rows of data and has a size of 9.05 Gb on disk. It is made up of country level datasets of daily time series data relative to covid 19 globally.

5.3.2 RapidMiner Experiments

RapidMiner is a tool with several versions such as RapidMiner Studio, RapidMiner Radoop, RapidMiner Server and RapidMiner Cloud. For this research, the RapidMiner Studio version was used since it is compatible with desktop computing and available on an educational license. To deal with missing values, several features were available in terms of filling data gaps, imputing missing values and replacing missing values. As missing value imputations is highly cited in the ML literature, the current study investigated its implementation through RapidMiner Studio. The imputation process accepts an 'exampleset', which is a dataset of values containing the raw data and returns a dataset with imputed values. Other processes, such as replacing missing values, can support the imputation process when the latter does not provide satisfactory results.

The 'filling of data gaps' process available in RapidMiner Studio was less relevant for this current research study. With this process, missing ID attributes values were calculated based on the greatest common divisor of distances of consecutive IDs. The other attributes are filled with a null value. Likewise, the 'replace missing values' process of RapidMiner Studio was not considered applicable for this current research study; missing values replacement is not very accurate as missing values are updated by a specific replacement value such as a minimum, average or maximum value of a given attribute.

The first dataset was connected as a local data repository with RapidMiner Studio, and the statistics feature revealed completeness issues in terms of missing values. For example, an attribute named 'CCN', which was a unique identifier for health care facilities certified to participate in federal health care programs, was reported to have a staggering amount of 1009941 missing values. Other attributes, such as 'Speciality', reported fewer number of missing values of only 38633.

To tackle data completeness issues in the first dataset, the 'impute missing value' operator was connected to the local repository containing the dataset in RapidMiner Studio. Within this operator, there is a sub-process, which takes the repository as input and would apply the K-NN algorithm for value imputation. This K-NN model replaces missing values by using Euclidean distance relative to available data to 'guess' more precisely what the missing values could be.

The first experiment involved the selection of all attributes to be imputed across all tuples in the dataset. However, upon execution of this process, the computation period went for 24 hours without achieving any output. The process was arbitrarily terminated, and a subset of attributes was selected, filtering out the 'CCN' identity attribute and including only attributes (NPI, zip, Provider_type) which could be correlated with the 'hospital type' attribute. The latter is one attribute which displayed subsequent units of missing values. However, even reducing the dimensionality of the dataset set to only four attributes resulted in the imputation process running without any output, before being arbitrarily terminated too.

The author suspected that the number of rows or examples involved, i.e., over 1 million, was a challenge to the computation capability of the software using a local desktop processing capability system. This suspicion was confirmed when the examples were filtered, taking only examples from range 2000 to 2100. Hence, the imputation process was applied on only 100 observations, with four attributes involved, and where a K-NN algorithm was used. However, this resulted in an error message saying implicitly that the amount of memory available was not sufficient to run this process. For all these reasons discussed above, no further RapidMiner based experiments were performed upon the other two datasets. *The use of RapidMiner Studio was judged to be inadequate for working with Big Data.*

5.3.3 Python Experiments

Python is a well-known programming language, and it is used extensively within the data science community. It possesses some interesting libraries such as 'scikit-learn' and 'impyute' to help deal with data quality issues. Python scripts were implemented over the 'Vertex AI' component of Google Cloud Platform. The notebook created hosted python 3 environment, n1-standard-4 (4 vCPUs, 15 GB RAM) machine type and 100 GB data disk. The three benchmark datasets were uploaded on a bucket as part of the 'Could Storage' component of GCP.

The first application of python for the current research study was an exploratory data analysis, with the aim of detecting attributes having missing values and their quantity. This was performed by applying the following algorithm:

```
import numpy as np
import pandas as pd
train_df = pd.read_csv('EHR.csv',dtype={"ZIP": object})
total = train_df.isnull().sum()
```

print(total)

The result was as follows for the first dataset in Table 5.2:

Table 5.2: Exploratory data analysis results of first dataset

Attribute	***# of missing values***
NPI	*0*
CCN	*1009941*
Provider_Type	*0*
Business_State_Territory	*0*
ZIP	*40845*
Specialty	*39237*
Hospital_Type	*1009941*
Program_Type	*0*
Program_Year	*0*
Provider_Stage_Number	*0*
Payment_Year	*76340*
Attestation_Month	*0*
Attestation_Year	*0*
MU_Definition_2014	*846028*
Stage_2_Scheduled_2014	*23250*
EHR_Certification_Number	*0*
EHR_Product_CHP_Id	*0*
Vendor_Name	*0*
EHR_Product_Name	*0*
EHR_Product_Version	*0*
Product_Classification	*8069*
Product_Setting	*8069*
Product_Certification_Edition_Yr	*0*

For the second big dataset, the result of the attributes showing number of missing values following the exploratory data analysis was as follows in Table 5.3:

Table 5.3: Exploratory data analysis of second dataset

Attribute	***# of missing values***
Denominator	6657
Payment_footnote	12167
Value of care footnote	12161
Location	1412

For the third big dataset, the result of the attributes showing number of missing values following the exploratory data analysis was as follows in Table 5.4:

Table 5.4: Exploratory data analysis of third dataset

Attribute	*# of missing values*
Data value	150518
Confidence_Limit_Low	152062
Confidence_Limit_High	152062
Data_Value_Footnote_Symbol	898056
Data_Value_Footnote	898056
GeoLocation	1545

Hence, it was clear that there are varying amounts of missing values across different attributes for all datasets. To be able to correctly use ML to detect DQ issues, the first step is to examine the properties of the missing values. For example, with the first dataset, it can be inferred that there is a correlation between 'CCN' and 'Hospital_Type', and between 'Product Classification' and 'Product Setting', based upon similar amount of missing values. Following the experiments from the 'RapidMiner' based experiments and its highly computationally intensive imputation results, it is important to filter out features whose values should not be imputed or replaced. The aim is obviously to reduce the computational complexity of catering for missing values in a Big Data context with a huge amount of data. In the current research study, the author decided to perform feature selection, which is recommended when applying ML algorithms to deal with data imputation in Big Data context (Ezzine & benhlima, 2018). Thus, to demonstrate missing value imputation and outlier detection experiments, only selected attributes were used.

The following step in the use of ML algorithms was to apply predictive modelling for imputing missing values in selected features. The general logic or pseudocode applied to all datasets for this step was the following:

Call the feature/attribute where you have missing values as y.
Split data into sets with missing values and without missing values, name the missing set X_test and the one without missing values X_train and take y (variable or feature where there is missing values) off the second set, naming it y_train.
Use one of ML algorithm derived to predict y_pred.
Add it to X_test as your y_test column. Then combine sets together.

The GAIN implementation followed a slightly different logic as it does not split into sets of missing and non-missing values but creates and incrementally optimizes masks matrices to store knowledge of missing values through application of loss functions.

Noise/outlier detection

This research study focused on detecting human or mechanically induced errors as part of the considered dataset. To know whether the three CSV real-world datasets used for the experiments faced inaccuracy issues, simple statistical analysis upon the datasets using count, mean, standard deviations, frequency, minimum and maximum values per attribute were carried out. This revealed with better clarity the accuracy problem/s which certain attribute/s might be facing. This knowledge needs to be coupled with the general context of use of the dataset to correctly discriminate between errors and acceptable extraordinary values. The result of an extreme value analysis upon the first big dataset, applicable only on numerical data, for detecting inaccuracy issues were as follows:

Table 5.5: Extreme value analysis results for first dataset

Feature	% of Outlier detected
CCN	0
ZIP	9
Program_Year	0
Payment_Year	0
Attestation_Month	10.2
Attestation_Year	0
MU_Definition_2014	0
Stage_2_Scheduled_2014	0
Product_Certification_Edition_Yr	0

The above clearly demonstrate issues with only ‘ZIP’ and ‘Attestation_Month’ features. The same process was applied for the other two datasets, with the following results only for attributes showing some quantity of outliers:

Table 5.6: Extreme value analysis results for second dataset

Feature	% of Outlier detected
Denominator	4
Payment_footnote	4

Table 5.7: Extreme value analysis results for third dataset

Feature	% of Outlier detected
Sample_size	12
Display order	16
LocationID	1

In the field of outlier detection, the discovery of outliers can trigger different possibilities ranging from automatic deletion of outlier values, replacement of outlier values and using human judgement for final decision upon outlier treatment and correction. As health data is highly sensitive and critical, the current study recommends use of human judgement for further actions regarding suspected outliers..

5.3.4 Vertex AI and BigQuery experiments on GCP

Vertext AI is one component available via GCP which would allow working with custom built python based ML algorithms which follow the logic described in section 5.3.3. A 'workbench' was created which provided an equivalent of a virtual python 3 based environment with the following main hardware specifications: machine type of 4 vCPUs and 15 GB RAM, data disk of 100 GB. The same ML scripts which had already been used on local desktop computing were run via Jupyter notebooks. The three CSV datasets were uploaded on a bucket as part of GCP Cloud storage and the only changes to the python scripts made were codes relative to connection with those datasets. The immediate advantage found was that certain ML algorithms such as isotonic regression implementation and CBMS which crashed upon local desktop computing conditions could be executed via this configuration.

Applying machine learning models on BigQuery datasets can be carried out via BigQuery ML via SQL queries. The limitation relative to this research study is that tailor made ML algorithms cannot be applied via this method and only ML algorithms which is supported by BigQuery ML were included as part of the BigQuery ML experiments. Hence, the algorithms which were included were Linear Regression. To determine the amount of missing values present as part of the 'snowfall_mm' data attribute of the 'covid_open_data' BigQuery public dataset, a simple SQL query was used and resulted into 15938434 missing values, hence confirming some data incompleteness issue as part of this attribute.

5.3.5 Experiments implementations

The following highlights the different experiments carried out. For imputation experiments, the 'Payment_Year' attribute was selected for first big dataset, 'Payment_footnote' attribute for the second big dataset and 'Confidence_limit_Low' for the third dataset. Concerning the BigQuery dataset, imputation was performed upon 'snowfall_mm' attribute which was of a numerical type. Whereas for detection of outliers for text values, experiments were performed upon 'Speciality' and 'Program_Type' attributes for the first big dataset, 'State' attribute for the second big dataset and 'Class' attribute for the third big dataset.

Bayesian isotonic and linear regression algorithms

From the details given in the research study consulted (Ahmed et al., 2016), it was not possible to have a complete source code, and which could guarantee correct replication of the Bayesian isotonic regression implementation. There were also issues in terms of the need to have some 'historic set of values' which should facilitate training. Hence, a decision was made to implement a very close alternative in terms of an 'isotonic regression' algorithm. However, this algorithm was found to be inadequate for Big Data as it cannot cater for the volume of data, but could be executed on the GCP platform. The results of imputation with this algorithm are detailed in section 5.4.1. As regression algorithms were commonly cited in existing research studies (Yu et al., 2017; Ahmed and Soomrani, 2016), linear regression was also implemented as it is described as quite close to isotonic regression algorithms (Ahmed and Soomrani, 2016). The linear model class from sklearn library in python 2.7 and python 3.8 were implemented both using local desktop computing conditions and on Vertex AI and it performed slightly better than isotonic regression algorithm. Furthermore, the linear regression algorithm was also implemented to impute missing values as part of the BigQuery experiments.

SRS-l_p

Following the guidelines of the original research study (Chen et al., 2018) detailing SRS-l_p, it was again found impossible to have a perfect replication due to missing details of the implementation. However, according to the algorithm documentation, the main functions of the algorithm is based upon the application of sparsity and regularization functions. To simulate these main functions this, an 'SDRegressor' class of sklearn library in python 2.7 and python 3.8 was implemented, with different parameters fine-tuned to have an emulation of SRS-l_p.This script was also experimented upon the Vertex AI environment.

Cluster-Based Best Match Scanning (CBMS)

This algorithm was again not properly explained as part of the original research study (Yu et al., 2017), but the logic understood from the algorithm implies the need to have regression of missing values performed upon clusters of data. Hence, k-means algorithm was applied for clustering and KNN for regression. The number of clusters, denoted by k, was set to 50 to have a better clustering of data points. Then, each cluster was split into 70% as training and 30% test set. Unfortunately, KNN algorithm cannot accommodate more than 1 million rows of values with the local computer experiments, but executed perfectly when the algorithm was executed over the GCP notebook.

Generative Adversarial Imputation Nets

Yoon et al. (2018) made some source code available via 'github'. One of the practical lessons learnt through these experiments was that different available source codes and python libraries required different versions of python installations and libraries version. Hence, another virtual environment had to be setup to execute the updated source codes of GAIN obtained to fit the current experiments. Some further data engineering was performed upon the selected features across the three datasets involved where the missing values were provided with the number zero. The different models were trained for 1000, 5000 and 10000 epochs and with the recommended parameters by Yoon et al. (2018) except that the missing rate simulation was cancelled.

Clustering combined with TF-IDF

Detecting errors in text data can be achieved using ML clustering such as *k-means*. However, as *k-means*, or any other clustering algorithm, cannot be applied on text data directly, there is the need to convert the text data into numerical data. During this conversion, each word is assigned a weight approximating its importance in a group of documents. For the current experiment, as there were no typographical or grammatical errors present as part of the attributes containing text data, errors were artificially introduced as part of two attributes in the first big dataset. Then, the "term frequency-inverse document frequency" (TF-IDF) algorithm was applied upon each attribute to produce a weight for each word. As there were many values that were repeating within an attribute, the *tf-idf* of these values were similar.

Following the transformation of text values into a series of *tf-idf* values, *k-means* algorithm was applied with the creation of only one cluster. The least important values were easily identified.

For example, when applying this method on the 'Program_Type' feature, the experimental algorithm output were "***Cluster 0: medicare medicaid hegfgf medigfgf".*** The same process was applied in the second dataset with the attribute 'State', and in the third dataset with the attribute 'Class', with comparable results as the one described with the attribute 'Program Type'. The last two values were artificially induced errors. Thus, with the application of the *k-means* algorithm with a single cluster upon the 'tf-idf' equivalents of words in a dataset, errors are normally outputted amongst the last in a cluster. Afterwards, human intervention is necessary to ascertain whether those last values are valid one or errors.

Neural networks

Neural networks was highlighted as one possible algorithm which might be used, with knowledge bases, to detect typos in text data. An experiment was carried out with an implementation of one the classical type of neural network, in terms of a Multi-Level Perceptron (MLP algorithm). It was carried out on the third real world CSV based dataset upon the 'Class' data attribute. Artificial errors were inserted as part of this attribute and these errors were listed at the end of a cluster by using clustering combined with TF-IDF. For the MLP implementation, the values were first converted with the doc2vec to convert the text into a vector space model. The vector size was assigned to 100 as the minimum threshold as the algorithm was highly computationally intensive. The second step was to apply an auto-encoder network in the form of the MLPRegressor library in python. The third step was to calculate similarity measure using a cosine similarity calculation. These steps are part of classical steps discussed as part of outlier detection on text data with the use of neural networks (Nag, 2019). This experiment was carried out over the vertex AI platform but took around 2 hours to provide results. Unfortunately, the outliers detected by this algorithm were not the artificial errors inserted in the data attribute. Hence, given this non detection of artificial errors and the relatively lengthy execution time of this algorithm, it is not recommended as adequate for detection of text outliers for EHR Big Data.

5.4 FINDINGS

The following table summarizes findings made after attempting to implement the chosen ML algorithms in Big Data for the healthcare industry. In many existing research studies, there is no implementation details provided, and therefore exact replication was not possible. However, the objectives of the ML algorithms were simulated in the performed experiments, as listed below.

Table 5.8: Summary of findings

ML Algorithm	Findings following implementation
Bayesian isotonic regression	Implementation exists only for isotonic regression, but required substantial computing resources and worked only with the GCP experiments; hence, had to implement a close alternative in the form of a linear regression algorithm. (https://scikit-learn.org/0.19/modules/generated/sklearn.isotonic.IsotonicRegression.html#sklearn.isotonic.IsotonicRegression)
Linear Regression	As regression algorithms are widely cited throughout literature, its most popular implementation was also experimented with.
l_p-norm regularisation (SRS-l_p)	SDRegressor library of the linear model was implemented. This model allows application of sparsity through the 'penalty' and 'l1_ratio' parameters and parameter 'alpha' which allows regularization. Those two parameters are the foundational building blocks of the SRS-l_p model, and therefore is deemed to have been successfully replicated. (https://scikit-learn.org/0.19/modules/generated/sklearn.linear_model.SGDRegressor.html#sklearn.linear_model.SGDRegressor)
Cluster-Based Best Match Scanning (CBMS)	Implemented CBMS by using Lyold's logic for k-means clustering and Pearson Correlation for the KNN regression. This implementation could only be experimented with the GCP platform.
Clustering combined with feature selection and imputation	Overlaps partly with the CBMS algorithm in the sense that this method proposes 3 phases; clustering, imputation and feature selection. The imputation phase was applied with 2 algorithms, namely KNN and Multivariate Imputation by Chained Equations (MICE). The CRI version with KNN is the equivalent of the CBMS,. As MICE is not adequately supported, the CRI algorithm was not considered as a potential ML support for DQ activities.
Generative Adversarial Imputation Nets (GAIN)	This source code obtained from Yoon et al. (2018) was modified to accommodate features from the datasets in the experiments. It must be noted that null values had to be converted to zeros beforehand else the numpy library would not have supported executing this algorithm. Even if GAIN proved highly plausible in terms of missing values imputation, it confirms that GANs based algorithms suffer from a relatively higher time complexity of execution.
Extreme value analysis	This statistical algorithm was applied on the features containing numerical data and found a certain percentage of outliers. However, the final vetting whether those outliers are errors should be based upon human judgement.
k-means algorithm combined with TF-IDF	TF-IDF was applied upon the text data of two attributes. As no errors were visible, artificial errors were introduced, and following the application of k-means for only one cluster, it is noted that errors, which were less and therefore rarer in the attribute, were listed last in the cluster. Subsequently, a human expert needs to ascertain whether those last values are errors or acceptable outliers. A pitfall of this method is the fact that there is the implicit assumption that values that are errors are rarer, and therefore would have less *tf-idf* weights. This assumption holds true in most cases, as if ever there are cases where there

	are more errors compared to accurate values in an attribute, this will be an indication of a systemic error as part of the data capture process. Solving this type of systemic problem is beyond the scope of this research, which focuses upon data driven rather process related issues in Big Data for the health industry. Furthermore, clustering-based technique does not require any additional external dictionary, which is an advantage as the technique does not depend upon the effectiveness of the dictionary.

An important piece of knowledge gained from this research study is that for the use of ML in the context of Big Data quality, the tool and processing power used for implementation certainly has a considerable impact. Highly efficient tools such as RapidMiner might not work in a certain version and might mandate the use of cloud-based versions. Similarly, many existing algorithms failed because of data structure or memory overload issues on local desktop computing but succeeded when experiments were executed via the GCP platform.

The different implementations made for this current study are provided in Appendix 1 and 2.

5.4.1 Evaluation of considered algorithms

Imputation

A crucial element to take into account for evaluating the accuracy and performance of algorithms implemented in the course of this research objective is that because of the use of real-world datasets, there is a lack of absolute certainty about what should be final corrected values, known as 'ground truth' samples. This lack of adequate test data for evaluation has been highlighted as one of the current challenging research aspects involved in the data and information quality domains (Becker et a., 2017). Some evaluation algorithms need the presence of ‘truth’ samples. Examples of such evaluation algorithms are ‘mean absolute errors’ or ‘precision score’. Hence, those algorithms had only been applied for performance evaluation whenever there was no technical possibility such as evaluating on the BigQuery ML component of GCP.

There are four types of imputation accuracy, as follows (Pasteels, 2013):

(i) Predictive accuracy or effectiveness: maximum preservation of true values (of each imputed value); (ii) Ranking accuracy: maximum preservation of true ordering (ranks) relationship in imputed values; (iii) Distributional accuracy: maximum preservation of the distributions of true values; and (iv) Global estimation accuracy: maximum preservation of analytic results and conclusions.

Due to the inherent lack of certainty relative to the correctness of values in a real-life Big Data scenario in the health industry, the global estimation accuracy imputation evaluation method was chosen as it does not directly rely upon knowledge of true values. In this case, the *plausibility* of imputed values could be used as another evaluation criteria for algorithms concerned with imputation. This is undertaken largely with the use of statistical data editing, but also with *outlier detection* (Sporleder et al., 2006). 'Leaving-one-out' approach to evaluating imputation method is typically used (Pasteels, 2013), but as it is reported to be a very time-consuming process, it was assumed not to be applicable for Big Data. This current research adopted the outlier value detection as a proper imputation evaluation method, since if ever some level of inaccuracy is induced after imputation, it will be detected by this evaluation approach. The actual outlier value detection was carried out using the z-score, which works perfectly to detect outliers amongst numerical data. If ever there are text data involved for imputation of missing values, the 'tf-idf' equivalent weights of the imputed replacement can be used. Methods of evaluation involving deleting some existing values from the dataset cannot guarantee accuracy of imputation evaluation as there is no certainty that the deleted values were accurate originally or possessed other DQ issues. Thus, if an imputation algorithm predicts another value other than the deleted one, it is impossible to posit whether the imputed value is correct or not in cases where 'truth' samples are not available.

The following tables provides a summary of differences in the number of outliers detected in the original datasets (OD) and datasets with imputations performed on a local desktop computer (IDl) and datasets with imputations performed over GCP (IDg). The logic is that if ever there is an almost similar number of outliers between OD and ID, then the imputed values are considered plausible. The columns Tl and Tg denote execution time in seconds. The following acronyms were used for the different algorithms: Linear Regression(LR), Isotonic Regression(IR), Lp-norm regularisation (SRS-l_p), Generative Adversarial Imputation Networks (GAIN) and Clustering combined with regression (CBMS).

Table 5.9: Plausibility of algorithms for first dataset(EHR.csv)

Algorithm	OD	IDl	Tl	IDg	Tg	Conclusion
LR	0	0	8.2	0	1.2	As no outliers have been introduced, this method is deemed to be plausible.
IR	NA	NA	NA	0	1.2	With GCP based experiments, use of IR is plausible.
CBMS	NA	NA	NA	680	1.6	A small amount of outliers detected, hence relatively not plausible.

(SRS-l_p)	0	76340	10.8	76340	1.3	As all the imputed values seems to have become outliers, this algorithm is deemed not plausible.
GAIN	0	0	240	0	85	As no outliers have been introduced, this method is deemed to be plausible

Table 5.10: Plausibility of algorithms for second dataset(pvch.csv)

Algorithm	OD	IDl	Tl	IDg	Tg	Conclusion
LR	488	488	0.06	488	0.02	No difference in outlier amount, therefore plausible.
IR	NA	NA	NA	488	0.02	Same as LR but with computing platform constraints
CBMS	0	82	0.01	46	0.01	Less outliers compared to Linear regression but relatively not plausible.
(SRS-l_p)	0	0	0.11	0	0.02	No outlier values were detected, this algorithm is deemed more plausible for smaller datasets.
GAIN	488	488	12	488	3.5	No difference in the amount of outliers, but with a greater time complexity compared to other algorithms.

Table 5.11: Plausibility of algorithms for third dataset(BRFSS.csv)

Algorithm	OD	IDl	Tl	IDg	Tg	Conclusion
LR	0	0	3.15	0	1.21	This algorithm is deemed to be plausible.
IR	NA	NA	NA	0	1.27	This algorithm is deemed to be plausible, but slightly worse than LR
(SRS-l_p)	0	26837	3.46	26837	1.21	As 18% of the imputed values seems to have become outliers, this algorithm does not show a relatively good level of plausibility.
CBMS	NA	NA	NA	2300	4	Some small level of outliers detected, therefore not plausible.
GAIN	0	0	220	0	128	This algorithm is deemed to be plausible.

Considering the first and third big dataset which involved more than 1 million rows of data, Linear Regression, Isotonic Regression and GAIN were the most plausible imputation algorithms with zero outliers detected following imputations. However, there is a major difference between the time complexity of those algorithms, with linear regression more adequate in terms of computational complexity. The difference in time complexity between Linear and Isotonic Regression algorithms is almost the same, but Isotonic regression could only be applied with far more computing resources. For those two datasets, SRS-l_p showed a high amount of outliers, with all imputations flagged as outliers for the first dataset and 18% for the third dataset. Concerning all the three datasets, CBMS was replicated with a relatively low amount of outliers detected. However, for the second dataset, SRS-l_p was as plausible as Linear Regression as it showed no outliers whereas GAIN again demonstrated a relatively lengthy execution time. The results of experiments on the third dataset are similar to the first dataset. It is quite clear that imputation techniques supported by ML is impacted by the volume of a dataset and the

computing power involved, and that for big datasets (more than 1 million rows of data), Linear Regression algorithms tend to be more plausible considering its high level of plausibility which is similar to GAIN but linear regression is far better in terms of the time complexity criteria relative to GAIN. This can be explained due to the fact that the imputation of missing values is based upon knowledge derived from non-missing data and regression algorithms typically performs well in this mode (Ahmed et al., 2016). The efficiency of regression algorithms for imputation of missing values is recognised in the area, specially upon small datasets as attested by Tran et al.(2015) and Lu et al. (2021).

As for the BigQuery public dataset experiments evaluation, the only imputation algorithm which was included for the experiments from the literature review and which could be applicable upon BigQuery ML was Linear Regression. A model was created to predict the missing values for the "snowfall_mm' data attribute. The model evaluation was carried using the following syntax "*SELECT * FROM ML.EVALUATE(MODEL `covid_open_data.penguins_model`)*" where the model name was 'penguins_model'. The mean absolute error of this evaluation was 156.23. For the sake of comparison, another regression algorithm available via BigQuery ML was chosen as classification algorithms could not be applied upon the data attribute which had more than 50 class labels. The other algorithm was a deep neural network regressor (DNNr) algorithm, given that GAIN is a variation of a neural network and GAIN was recommended to be part of the experiments. Upon evaluating the missing value predictions from the DNNr model, its mean absolute value was found to be 224.48. Hence, upon comparing linear regression with deep neural networks for data imputation in the context of this experiment, linear regression proved to be imputing missing values better.

Outlier detection evaluation

Accuracy DQD issues detection of numerical data was based upon the use of z-score, which is a well-known function for outlier detection. As there was no certainty over which values are errors without a 'truth' sample, there was the need to induce artificial errors. Evaluating z-score will not bring any new contribution to knowledge, hence this was not performed as part of this research.

To detect potential outliers as part of text data around 20 errors were introduced in the 'Program Type' attribute to evaluate the error detection of text data for the first big dataset. 15 of those errors were just repeating dummy words such as ' hegfgf', and 4 were only single alphabets, such as 'b'. All the dummy words were highlighted as part of the clustering process, but not the ones

made up of single alphabet . This could be explained by the 'stop words' parameter set to English in the TF-IDF ‘vectorizer’ process. In any case, the detection of dummy words would result in very high precision and recall benchmarks for inaccurate terms detection following the clustering. Another experiment was carried out using the ' Specialty' attribute, with the raw original data for the first benchmark dataset. The clustering process highlighted 75 distinct terms, and upon manual inspection, they were all correct English terms. Hence, they were assumed to be accurate data. With the third benchmark dataset, dummy terms were inserted in the ‘Class’ data attribute and the clustering process listed those dummy words at the end of a cluster, same as the experiment with the first dataset. Those experiments clearly demonstrate that the use of clustering algorithms combined with TF-IDF group text outliers together, but also group correct text data values in other clusters. The efficiency of using this ML based technique for text data outliers is therefore clearly established.

The final process of ascertaining whether a term is an error would rely upon human judgement as it is extremely challenging to distinguish correct and incorrect data values automatically, without a training process using external knowledge through dictionary of terms which had been explained to be inappropriate for Big Data in the accuracy DQD discussions of this study. The use of human judgement to detect some types of errors out of datasets had also been proposed as part of a technique called ‘ADQuate’ which also discussed about use of ML algorithms to detect constraints violation in data mining processes (Homayouni et al., 2019).

5.5 CONCLUSION

The overall aim of this chapter was to investigate the possibilities of using ML algorithms to improve data quality operations, more specifically concerned with data completeness and accuracy. The rationale for the use of ML algorithms stemmed from the characteristics of Big Data, namely volume, variety and velocity, which challenge data quality methods applicable in a non-Big Data context. This chapter focused on the research objective of investigating the use of ML algorithms for detecting dirty data as part of a Big Data representation of EHR data. The findings of this chapter are intended to form part of a more holistic data quality methodological approach aimed at improving Big Data quality for the health industry.

The literature review concluded that there were some previous uses of ML relative to the completeness DQD; however, most of the existing research studies involved imputation of missing values in the most effective way through ML, and not specifically the use of ML algorithms to **detect** missing values. Further investigation showed that the detection of missing values in a dataset is a straightforward process with tools such as RapidMiner Studio or programming languages such as python. Disguised Missing values were not considered for this research study, as they might also be treated as outliers or inliers (Qahtan et al., 2018).

Consequently, the research objective diverted from its original aim of only detecting missing values and carried out experiments to determine which ML algorithms could be most effective for imputation of missing values within EHR Big Data. The experiments were conducted on three real world CSV based EHR datasets and 1 BigQuery dataset. The results concluded that the use of linear regression was best in terms of the accuracy of data values imputed and computation time of imputations both with experiments performed on local computing settings and over a cloud-based architecture. This confirms the findings from several other research studies in the area of missing data imputation which detailed how regression based algorithms performs very well (Tran et al.,2015; Lu et al.,2021). Furthermore, the techniques described in a US patent file also proposed Piecewise Linear Regression imputation model for both small and large/distributed data sources for continuous non categorical values (Chu et al.,2016). One challenge related to the evaluation of the experiments because the latter involved real-world datasets, and therefore, the correct data values were unknown. Due to this fact, the plausibility evaluation measure was applied. On the other hand, it is a reasonable assumption that in a Big Data context, knowledge

of correct data values would be limited, and hence, the plausibility evaluation technique would fit well.

Regarding the use of ML algorithms to detect data inaccuracy, the first conclusion is that there is not a unique ML algorithm that will be able to cater for all types of data inaccuracy issues. Inaccuracy issues, focused on outlier values, as part of numerical data can be detected with non-ML algorithms such as the use of the statistical algorithm known as 'z-score'. ML algorithms can however be useful to detect outlier text data where a transformation of the text data into TF-IDF scores is required first, and then k-means (a clustering-based ML algorithm) was applied. The results showed that some artificially induced text errors were detected in this way. With the use of clustering algorithms, potential suspected inaccurate data will be more easily grouped together and will therefore ease the process of inaccurate data detection. Hence, this research concludes that human expertise is needed to validate potential errors which had been highlighted by a clustering algorithm. Thus, a semi-automated approach is advised as part of data inaccuracy detection systems for Big Data in the health industry. The semi-automated approach will operate as such (1) use clustering algorithms to group and highlight outliers (2) ask human experts to ascertain whether highlighted outliers are errors or acceptable values.

Overall, the experiments carried out as part of this research proved that it is very difficult to have an umbrella ML algorithm category capable of dealing with both types of DQ issues. In the case of missing value imputations, regression-based algorithms tend to be more effective, both in terms of plausibility and computation time. Whereas in the case of data inaccuracy issues, specifically for text data, clustering based ML algorithms are more effective. Furthermore, detection of missing values and detection of outliers amongst numerical data do not specifically require the application of ML algorithms and can be performed adequately using known statistical functions. Concerning the superiority that the unsupervised category of ML algorithms assumed in the research methodology chapter, it has not been proven. Even if clustering algorithm, which is an example of unsupervised learning, is more adequate for text-based outlier detection, this chapter has shown that supervised learning techniques are also applicable to some level. Those techniques do not necessarily need to base themselves upon predefined manually annotated labels nor reference datasets to detect some types of DQ issues.

Therefore, a system aimed at improving DQ for EHR Big Data will need to devise a hybrid solution, mixing regression and clustering-based ML algorithms with statistical functions for the

detection of data incompleteness and inaccuracy issues. This solution should also be semi-automated, as the involvement of human expertise is deemed to be essential for detecting inaccuracy errors. The technology used to develop the solution also has a heavy impact on the effectiveness of the solution. It is recommended that technologies allowing the application of cloud-based architecture to should be used to implement the ML algorithms discussed in this chapter.

Chapter 6: Investigating Data Repair steps for Big Data in Health industry

6.1 Introduction

In this chapter, the process of improvement of data quality is investigated. The goal is to understand how best data repairs or transformation should be performed, in terms of steps and activities, in the context of EHR Big Data. The focus of data repairs follows the results coming from both Chapters 4 and 5. Data repairs for the completeness DQD were performed by data imputation algorithm whereas data repairs for the accuracy DQD concentrated on how to best deal with outlier values. The hypothesis followed throughout this research study is that a completely automated data cleansing approach is not the most suitable, hence, the need for a minimum amount of human intervention. Automated data repairs risk to produce systematic and random errors varying with tools and users (Sukumar et al., 2015). The amount and nature of this human intervention is further investigated to meet the research objective in this chapter.

A new prototype of data repair algorithm was developed and compared to as part of some of the experiments of this chapter. The algorithms intended to be part of the experiments are existing Big Data quality tools such as BigDansing, BayesWipe, already detailed in the literature review chapter of this thesis and other tools discussed as part of academic literature not yet discussed in the thesis report, such as Cleanix, ActiveClean and HoloClean, amongst others. The new prototype was developed based on the knowledge of the use of ML to help improve data inaccuracy and data incompleteness in Chapter 5. With this comparison, the aim was to assess efficiency and effectiveness of various methods, steps and activities used for data repair. Furthermore, a survey of existing research studies discussing data repair tools, algorithms and methods was also performed in order to broaden knowledge about steps and activities related to data repair for Big Data.

Most of the different algorithms and tools mentioned in this chapter had been applied on the CSV real world datasets described in the previous chapter in Section 5.4.1 as part of experiments. Wherever some algorithms and tools were not included as part of experiments, it was due to lack of availability or lack of their implementation details. Yet some methods of those algorithms and

tools were still important for informing the steps relative to data improvement of a proposed data quality methodological approach and therefore included for discussion and consideration.

This chapter investigated the performance of some existing algorithms and tools mentioned during the literature review. As data repairs for Big Data is a relatively new research domain and hence few algorithms and tools were found from the literature review, other algorithms and tools were investigated and accessed through search engines. Some might not be typically Big Data cleansing tools, but they claim to be able to repair data in the context of Big Data. Only the free and available versions of the algorithms and tools are surveyed due to funding limitations, but the total number of algorithms/tools considered is judged adequate to carry out the controlled experiments to get an idea of the techniques and methods active in the data cleansing domain.

Not all data cleansing algorithms/tools found were considered adequate for the experiments. E.g, 'Informatica Data Quality' caters for the whole data quality lifecycle and therefore does not focus specifically upon the data repair or transformation process. They aim to reduce data entry errors by activating automation and mistake proofing mechanisms. Therefore, these types of tools are not investigated as part of this chapter.

This chapter is divided into the following sections: Section 6.2 gives an in-depth discussion about tools, steps and techniques that are adequate for data repair, in the context of Big Data; Section 6.3 explains the experiments carried out, and the knowledge gained from them and Section 6.4 discusses the principal steps, activities and characteristics for data repair, which informs a proposed data quality methodological approach for Big Data in the health industry.

6.2 Review of existing data repair algorithms and tools

Boostclean

Machine Learning (ML) was used to support data cleansing with an algorithm known as 'BoostClean', where a small clean training data set can be used for learning data repair rules (Krishnan et al., 2017). The evaluation criteria of this tool included, firstly, the performance of the imputation function for missing values in terms of the number of correct imputations and computational complexity, and secondly, the amount of inaccuracy provided by potential outliers mostly for text data. Therefore, the evaluation of 'BoostClean' is for similar data quality issues as was discovered in Chapter 5. Krishnan et al. (2017) explained that using scripts already developed by software engineers is not a good data repair approach, since predictive applications

deal with unknown data errors, which those pre-written scripts are not apt to cater for. Hence, use of ML can be more convenient for data repairs for Big Data, as there is a need to have algorithms which needs to learn from data characteristics to improve themselves and more unknown errors are expected with Big Data. 'BoostClean' performs automated data cleansing by relying upon statistical boosting which uses the best ensemble of operations from a library of ML algorithms. To achieve this, there is the need for a gold standard dataset which provides correct labels for the training dataset. As explained before in this current research study, the gold standard dataset does not tally properly with the concept of Big Data, but as discovered in Chapter 5, a supervised learning-based approach may not necessarily need a gold standard dataset but can use existing data to learn and generate models.

REDs

Another recent study, which proposed a new data-cleaning pipeline prototype called REDs, performed a limited comparison of some data cleaning algorithms and tools (Mahdavi et al., 2019). This pipeline was made up of three main parts: 1) data profiling, 2) detecting errors and 3) generating datasets cleaning workflows. These parts are to some extent like the parts discussed in the data quality methodological approach explained in Chapter 1 of this current research. Another important similarity with the current research study is that the data-cleaning pipeline makes use of both ML algorithms and off-the-shelf software. However, this data cleaning prototype is not focused on Big Data. It makes an interesting use of an ensemble learning stacking technique to improve error detection. This **ensemble learning method** briefly operates as thus: 1) there is the training of first level classifiers on a common dataset, namely a neural network, a decision tree and naïve Bayes classifiers. The output from those first level classifiers, typically known as a model, is then used to train a meta-classifier using logistic regression. In terms of the performance with the use of the harmonic mean benchmark only, NADEEF(FD) reports a ratio of 27%, WRANGLER 21%, outlier detection using Gaussian algorithm only 23%. Even if this ensemble learning-based stacking method shows an impressive amount of recall benchmark, in terms of 91%, there is an issue as 1% of dataset size must be labelled for it to work. This **labelling process** could be very ineffective and unrealistic in a Big Data context. Unfortunately, there are no discussions of comparisons involving the individual ML algorithms cited above by Mahdavi et al. (2019), which would have allowed a more adequate comparison for the current research objective.

BayesWipe

The latest version of 'BayesWipe' was downloaded from 'http://bayeswipe.sushovan.de/'. BayesWipe is explained to be applicable upon any comma-separated file. It is completely automated and based on Bayes networks. The characteristics, including use of statistical/machine learning algorithms, of this tool are discussed in the literature review of this thesis and also in more depth as part of experiments discussed in section 6.3 of this chapter. BayesWipe was very important in forming the foundational assumptions and hypothesis driving this research, as it was one of the first algorithm found which claimed to perform data repairs on Big Data systems. Hence, it was imperative to include it in experiments on the benchmark datasets.

BigDansing approach

To the best of the author's knowledge, there is no readily available algorithm for the 'BigDansing' approach. Details regarding the main methods which BigDansing makes use of had already been discussed as part of section 2.2.9 of the literature review of this thesis. In general, BigDansing relies upon user defined rules rather than machine learning to perform data cleansing. However, BigDansing is one of the few approaches claiming at undertaking data cleansing for Big Data, and therefore again was thought to be a very good candidate to include in the experiments. There were pseudocodes discussing the implementation of some of the modules in the related research study (Khayyat, et al., 2015). Unfortunately, there were many missing gaps in the description found in the research study to deploy the whole system,. This meant that there was no guarantee that a genuine version of BigDansing could be replicated for this comparison. Furthermore, the creation of user defined rules for data violations is not too realistic with real world datasets. Hence, the final decision was taken not to include BigDansing for the experiments during this section of the research.

ActiveClean

This is a progressive data repair algorithm, where a machine learning model is updated incrementally instead of re-training the whole model frequently. Thus, there is an interactive training of model-cleaning of dirty data iteration which is facilitated by this algorithm. One key component is a gradient function which should produce gradient of the loss. With convex loss models such as linear regression and SVM, the gradients of the loss are well-known, whereas with non-convex models, the gradient of the loss needs to be expressed programmatically

(Krishnan et al., 2015). An important point to consider is that using a mixture of clean and dirty data to train a model can lead to unreliable results for data cleansing (Krishnan et al., 2015). This is impactful for the current research study since many algorithms discussed in chapter 5 follow this principle. Even if 'ActiveClean' is reported to work well on small samples only, and therefore not on Big Data systems, its analysis provides some important ideas. Firstly, there is the mix of automated mechanisms of data cleansing with human judgement, that will specify the first model or set of data cleansing rules. Users are also allowed to train predictive models while progressively cleansing data. 'ActiveClean' has been used in experiments with a mixture of datasets with different types of data and returned highly satisfactory results (Krishnan et al., 2015). 'ActiveClean' also makes use of machine learning both to detect and to repair dirty data. Ultimately, the most important contribution of 'ActiveClean' is the progressive learning of the model, which lowers the cost of learning and is well suited in a dynamic data environment which might be expected from a Big Data system. To a lesser extent, there is also the knowledge of mixture of automated and user-based activities for the data cleansing process which might inform the proposed data quality methodological approach for the thesis.

SCARE

SCARE uses a data repairing approach based on improving imputed data given a data distribution and can be modelled using statistical machine learning techniques and likelihood methods (Yakout et al., 2013). It makes use of horizontal data partitioning, which allows different updates for the same record, and local predictions are ultimately combined to form a final prediction. SCARE not only caters for data imputation and data deduplication, but also for repairing of erroneous data with the use of machine learning. Therefore, SCARE caters for similar DQ issues compared to the current research study. Several challenges had been identified with the use of ML for data repairs (Yakout et al., 2013), such as:

1) There might be several dirty attributes within a record, and therefore, correlations between attributes might not be effective in learning clean data models.
2) The process of learning a model from a very large dataset is expensive, and the model itself may not hold in main memory. This limitation had been experienced in the current research study during the investigation of ML techniques to detect data quality issues in chapter 5 and highlights the importance of the computing platform to be used for DQ activities on Big Data.

SCARE can apply data imputation upon all data types and is not limited to numerical and categorical types. It also combines both a local and global view approach to data relationships, and therefore improves accuracy of data repairs.

It is unfortunate that this technique could not be replicated and benchmarked against the proposed prototype as it provided some very promising features. There is no algorithm or pseudocode discussed in the research study by Yakout et al. (2013), nor are they available via the web. However, there is the confirmation that application of ML algorithms upon a large volume of data can be a constraining factor. It encourages the deduction that data repairs for Big Data might only be carried out upon cloud-based or parallel processing-based platforms.

Cleanix

This is another algorithm cited in existing research studies which contains some important and connected characteristics with this current research study, but for which, there was no way of comparing with the proposed prototype, as no source or pseudo codes were available (Wang et al., 2014). The authors claim that 'Cleanix' can handle four types of data quality issues, including *abnormal value detection/correction* and *incomplete data filling.* These are the two types of dirty data issues that the current research study is also focusing upon. 'Cleanix' is developed with the 'Hyracks' execution engine, which is a data-parallel execution engine for Big Data computations. It allows as source, a stream of data, hence it appears very promising to handle the velocity aspect of Big Data. Thus, it should be theoretically applicable upon Big Data, even if user defined, instead of automated, data repair rules are applied. Unfortunately, there is a lack of discussions on the evaluation of the data repairs done. But 'Cleanix' is another indication that data repair for Big Data might require cloud or parallel computing-based platforms.

HoloClean

'HoloClean' focuses on data repair features only and does not possess error detection capabilities. It makes use of probabilistic models, just as 'BayesWipe', for correcting errors that stem from 'Denial Constraints' (DC) techniques. This tool has been compared against four other data cleansing tools (Rekatsinas et al., 2017) Even if they are not focused on Big Data, some of the datasets used in the evaluation can be considered as large in terms of volume. The other tools in the comparison were 'Holistic' (Chu et al., 2013), 'KATARA' (Chu et al., 2015) and 'SCARE' (Yakout et al., 2013). For datasets with low number of errors, 'HoloClean' achieved a high precision ratio of data repairs, whereas for datasets with a high percentage of errors, 'HoloClean'

achieved a recall ratio of 66.9%. The evaluation benchmark results of 'Holistic' were fair, but not as good as 'Holoclean' for both precision and recall. 'KATARA' obtained very high precision but limited recall. 'SCARE' performed well in datasets which it might learn from correct records but did not terminate processing for two datasets. On the negative side, the running time of 'HoloClean' was the worst globally amongst all the tools in the experiment (Rekatsinas et al., 2017). The use of DC rules constitutes another potential limitation of the techniques used by 'HoloClean' for Big Data use cases as this assumes a pre-defined knowledge of DQ issues.

Data Prep using Trifacta Wrangler

Trifacta Wrangler can deal with missing data and perform other data transformation/pre-processing tasks. It has been ranked as one of the best data cleansing software following a general search upon search engines, hence its consideration in this review. The search term used was 'Data cleansing software for Big Data'. The top-most link discussing data cleansing tools in general was "https://www.dsxhub.org/data-cleansing-top-of-the-best-tools-to-clean-up-your-data/". From this link, the off the shelfs tools discussed in this chapter were referenced. However, this tool cannot accept datasets with a size greater than 100MB, and therefore was initially considered not to be applicable for Big Data. However, with Google Cloud Platform (GCP), there is a tool known as 'Dataprep' which calls the Trifacta Wrangler software to perform data preprocessing tasks on data. Upon uploading our benchmark CSV datasets, it was possible to perform some basic data imputations where a user can specify rules which would replace missing values with a constant, such as the number '0' for integer based data. As discussed in chapter 5, imputation using a constant value creates the threat of inserting new errors in a dataset and is considered a reductive and overly simplistic method to deal with missing values. Hence, given the limited data cleansing features proposed by Trifacta Wrangler, it was not compared against the proposed prototype.

From the main existing data cleansing algorithms/tools investigated from existing research and analysis of off-the-shelfs software, the following important ideas was deducted:

1) most of the algorithms/tools are not apt to work with Big Data volume characteristics.
2) Most algorithms provide data profiling/exploratory analysis features.
3) Basic missing values replacement features are usually available. However, very little use of machine learning for imputation was found.

4) Many data repair features were usually provided, but these do not address data accuracy issues, but focused more on data correctness.
5) Existing algorithms/tools focus more on certain specific DQ activities such as cleaning of addresses, deduplication of customer records or email data cleansing. These reflect common and classical business needs, but DQ activities with use of Big Data might be broader and also different depending upon the context of application.
6) No actual feature to take care of data repairs after detecting outliers automatically.
7) A predefined knowledge of accuracy errors is required and must be inserted in the algorithms/tools in the form of rules or denial constraints. Other algorithms/tools make use of a golden dataset but none of these methods are very realistic for a Big Data context as they might be unknown or not available.

Furthermore, from a previous study reviewing existing data cleansing methods for Big Data (Ridzuan & Zainon, 2019), there are some interesting correspondences with review of existing literature above, as follows:

1) Only manual data cleansing is not appropriate with Big Data.
2) The complexity of data quality algorithms increases due to the inherent 3 V's characteristics of Big Data.
3) The volume of Big Data is an issue with data cleansing algorithms/tools operational upon non-Big Data scenarios.
4) Most existing algorithms/tools follow a constraint-based, or rule-based, approach to data cleansing. This may not capture different and changing types of errors, as expected in a Big Data scenario.
5) In the event existing data cleansing algorithms/tools correctly detect data quality errors, the quality of the data repair process is not well known and hence new errors may be introduced.
6) Human judgement might be mandatory to validate data repairs; however, human involvement in the data cleansing process should be minimized for cost and performance optimization. This could be implemented through crowd sourcing techniques such as the one adopted by the 'KATARA' system to minimize cost.

6.3 Experiments

The experiments involved in this chapter aims to compare data repair tools and algorithms as per the DQ issues highlighted in Chapter 5. As discussed in the Section 6.2, even if there are many discussed data cleansing algorithms/tools, there are very few of them that could be replicated for the data repair experiments. This is mostly due to practical issues such as improper implementations of the tools available or lack of precision in the pseudocodes discussed in available research studies. Thus, the same experiment methodology as adopted by Yakout et al. (2014) for their proposed 'SCARE' algorithm was followed where datasets used in the original studies were used if ever they could be considered as EHR datasets and were available for download or access. Secondly, one benchmark EHR CSV based dataset used for this thesis was also inserted as part of the Raha and Baran algorithm and compared with other algorithms.

The following details the experiments carried out:

Experiment 1:Data repair prototype

A prototype merging regression and clustering algorithms as detailed in Chapter 5 was created. This prototype performed data cleansing as it accounted for data imputations automatically and allowed human experts to classify outliers, ultimately performing updates and deletions wherever needed. Note that numerical data outliers are detected via classical outlier detection statistical algorithm such as 'z-score', and hence, as part of univariate data repair method, a cleaning parameter could have been set and therefore all detected outliers would have been removed. However, this method is deemed risky as the outliers had not been verified and confirmed to be errors. This is the main rationale why human judgement is recommended to take care of both numerical and text-based data outliers.

As the other algorithms could be experimented only with the pvch.csv dataset with acceptable computational complexity, the prototype was also evaluated with this dataset. The implementation and source code of the prototype is provided as part of Appendix 2 of this thesis. Some of the evaluation results were already obtained in Chapter 5 while others using metrics such as mean absolute error, precision, recall and f1 score were performed as part of the experiments for this chapter. As the data repair prototype makes use of linear regression for imputation of missing values, the mean absolute error (MAE) was used as it was a regression and not classification problem. The MAE score was unsurprisingly 0.0 as no outliers were detected

after imputation. Hence, the prototype provided excellent results in terms of the quality and time complexity of automatic data imputation.

Experiment 2: Rahan and Baran algorithms were cloned upon a virtual machine instance on GCP. The datasets used as part of the evaluation can be considered to be small datasets with the largest one having only around 200,000 tuples. The 'benchmark.py' script was run without the 'fast' parameter as per the reproducibility notes provided by the authors. Unfortunately, after 2 hours of execution, no actual results were obtained and the decision was taken to terminate the experiment in this mode as it was clear that the experiments in this mode was not applicable to Big Data. When the 'fast' parameter was included, the benchmarks ran for less than half an hour with the following main results in Figure 6.1:

Comparison with the stand-alone error detection tools. (Precision, recall, f1 score)

Approach	hospital	flights	beers	rayyan	movies_1
dBoost	0.54, 0.45, 0.49	0.73, 0.58, 0.65	0.59, 1.00, 0.75	0.15, 0.84, 0.25	0.25, 0.79, 0.38
NADEEF	0.05, 0.37, 0.09	0.42, 0.93, 0.58	0.13, 0.06, 0.08	0.30, 0.85, 0.44	1.00, 0.08, 0.16
KATARA	0.44, 0.11, 0.18	0.00, 0.00, 0.00	0.00, 0.00, 0.00	0.00, 0.00, 0.00	0.00, 0.00, 0.00
ActiveClean	0.02, 0.09, 0.04	0.30, 0.99, 0.46	0.16, 1.00, 0.28	0.09, 1.00, 0.16	0.06, 1.00, 0.12
Raha	0.99, 0.50, 0.67	0.89, 0.82, 0.85	1.00, 1.00, 1.00	0.88, 0.82, 0.85	0.84, 0.97, 0.90

Figure 6.1:Comparison of data repair algorithms

It was clear that Raha was better compared to NADEEF, KATARA and ActiveClean in all the datasets involved in the above experiment. The first column refers to precision scores, second column to recall score and third column to f1 score. Focusing upon the hospital dataset which was a small EHR dataset, KATARA was found to have a better precision compared to NADEEF and ActiveClean, even if this is not the case for the other datasets.

The next step was to perform those comparisons upon one of the benchmark CSV real world datasets discussed in chapter 5, namely the pvch.csv dataset. This dataset was chosen as it was the smallest amongst the three of the benchmark datasets, yet relatively bigger than the datasets used for the original Raha/Baran experiments. As such, the first step was to provide a clean copy with truth samples of data for this dataset. This presented a major risk as there was no way to know what should have been correct data items and hence this step was performed mainly by replacing null values by zero (0). As only a sample of only 20 values out of around 18000 would be taken for the learning process, this risk was deemed worth taking. As the learning process of the error detection strategies is automated, these algorithms tend to check associations and correlations amongst all the data attributes and hence becomes more computationally expensive

as the dataset grows in size, which is not so encouraging for Big Data contexts. It took approximatively 1 hour to create 1018 strategies for the pvch.csv dataset. The error detection performance was 0.98 for precision,1.00 for recall and 0.99 for f1 score whereas the error correction experiment crashed due to memory error.

The same dataset was executed by dBoost, KATARA and ActiveClean also. The results of this execution will allow comparing amongst those abovenamed algorithms, but definitely those algorithms face huge difficulties with the volume of data even when being executed via a cloud based platform. However, this part of the experiment was too computationally expensive in terms of memory use with 4 vCPUs and 15GB RAM. The platform was hence increased to 16 vCPUs and 60 GB RAM and finally, the following results were obtained after 1 hour of processing:

Approach	EHR
dBoost	0.21, 0.65, 0.31
NADEEF	0.00, 0.00, 0.00
KATARA	0.00, 0.00, 0.00
ActiveClean	0.04, 1.00, 0.08
Raha	1.00, 1.00, 1.00

Figure 6.2: Benchmarking of algorithms on pvch.csv

The results above give us a clearer indication of performances of some of the ML algorithms which might be used for data repairs in the context of EHR Big Data. NADEEF was eliminated early on during the experiments as there was the need to include user defined functions and patterns, which is something which is not realistic with Big Data. KATARA showed 0% for all three metrics which might be explained by the fact that this algorithm required an external reference set or dictionary in order to perform data repairs. The external reference set was based on Wikipedia pages in the above experiment and might not show many correspondences with the EHR data attribute values. Hence, the data repairs performed by KATARA were either totally wrong or could not be evaluated. In any case, it is quite clear that KATARA is not a ML algorithm which is adequate for EHR Big Data. ActiveClean denoted 4% level of precision but 100% recall, which might be explained by the fact that ActiveClean performs data repairs on a tuple level and not on instance level of a dataset. Upon evaluating upon tuple level metrics, the following results are denoted:

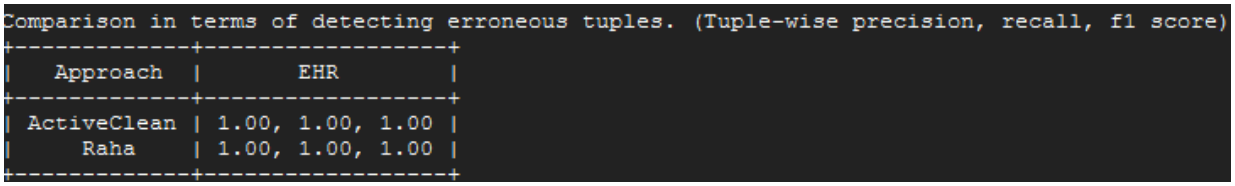

Figure 6.3: Evaluation using tuple level metrics

In this situation, both ActiveClean and Raha denote 100% for all the three metrics. This is to be taken with lots of caution and can be attributed to the fact that the 'clean.csv' version of the dataset containing the golden standard of data was not available and had been synthetically created solely for the purposes of the experiment.

Therefore, this experiment proves that:

1) Performing data repairs on Big Data is extremely challenging, even impossible for many algorithms such as NADEEF and KATARA. This confirms the current research study's initial evaluation not to include those algorithms for further experiments.
2) The bigger the dataset, the less responsive the data repair algorithms. For a relatively medium size datasets, it needed a very powerful cloud based computing platform on GCP to be able to terminate execution successfully but after more than one hour. Thus, for real time applications which need to work with high quality data, data repairs with Big Data can be a serious issue.
3) Dependence upon a golden standard dataset does not provide very trustworthy data repairs as indicated by the perfect scores scored by Raha and ActiveClean. It is more probable that the data repairs were erroneous but matched the overly simplistic version of the golden standard dataset used. Ultimately, data repairs on Big Data cannot depend upon algorithms which require a golden standard dataset or reference to external dictionaries.
4) Evaluation of data repair ML based algorithms should not be using metrics such as precision, recall and f1 score as the lack of knowledge of actual correct values from real world datasets result into non trustworthy evaluation results.

*Experiment 3***:** BayesWipe was downloaded and was executed after installing the Java Virtual machine, which was a pre-requisite to run the 'Banjo' library. In the first attempt, all the attributes from the first real-world benchmark dataset as per Chapter 5 was selected. However, the execution went on for several hours without finishing during the learning network phase. This

was already an indication that BayesWipe does not run satisfactorily with Big Data. However, as the prototype developed for this study was experimenting with specific attributes, some feature selection was performed upon the original dataset and only three attributes were kept. These attributes were 'Program_Type', 'Payment_Year' and 'Zip'. However, BayesWipe outputted an error relative to the fact that the software could not process a high score network during step 3. A further dimensionality reduction step was then carried out, reducing the number of tuples in the dataset to just 2500 compared to around 1 million. With this limited amount of data, BayesWipe was able to complete its data cleansing task. The first inspection of its output file revealed that it had replaced missing values by '0' as the only activity of data cleansing. Some artificial errors were introduced in the 'Program_Type' and 'Zip' attributes. The typographical errors introduced in 'Program_Type' was updated to an acceptable value, but values in the 'Zip' attribute were all wrongly updated to values with 1 digit. BayesWipe does not allow any input from a cloud-based data source and hence only local desktop computing experiments with it were possible. Following the results obtained from the experiments in the first dataset, it was not judged useful to carry out the experiments with the second and third datasets as discussed in Chapter 5. Therefore, 'BayesWipe' was judged inappropriate for Big Data, and not appropriate for data repairs even on a small dataset containing numerical values.

As discussed in Chapter 5, this thesis has taken a human-based intervention approach concerning outlier detection and repairs. For outlier detection, statistical methods were found to be adequate for numerical data whereas clustering algorithms could help to find text data outliers more quickly. One of the most recent proposals for data outlier repairs termed as DISC is to make use of distance based algorithms based on lower and upper bounds of values (Song et al.). The authors benchmarked DISC against algorithms such as HoloClean, Eracer and Holistic. Thus, the current thesis, being unable to include HoloClean as part of experiments, will use the work of Song et al. to discuss the performance of outlier repair algorithms for clustering applications. For the flight dataset containing around 2000000 tuples, the experiments showed that DISC had an f1-score of 0.75, ERACER 0.69, HoloClean 0.65 and Holistic 0.67. Hence, a statistical based algorithm such as DISC seems to perform better than a ML based algorithm such as HoloClean. However, the experiments by Song et al. focus mostly on turning outliers into inliers but might generate new errors as the evaluation focus upon potential clustering applications after the data had been repaired by the different algorithms mentioned in this paragraph. This potential limitation of DISC strengthens the approach of this current thesis in terms of relying upon human judgement to perform the final data repairs, specially focused upon outliers.

Table 6.1 summarizes the conclusions derived from the secondary data collection, research articles and web sites, and experiments carried out.

Table 6.1: Conclusions about data cleansing tools

Tool	Conclusions	Comparison
Custom prototype	Makes use of ML to cater for missing values and outliers; validation of actual noise performed by human experts.	Performance benchmarks detailed in chapter 5, but the cost involved with the human intervention grows with the number of outliers to validate.
BayesWipe	Advocates for both a data source and an error models based upon statistical processes.	Execution did not finish when executing whole dataset and terminated with error after dimensionality reduction. Hence, the conclusion is that BayesWipe is not apt for Big Data cleansing.
ActiveClean	Makes use of ML; progressively learns models.	compared through the Raha and Baran experiments
SCARE	Most applicable for data imputation based upon statistical learning of data distribution.	No existing implementation found.
HoloClean	Based upon machine learning with very good evaluation benchmarks.	Source code available but could not be executed either with python 2.7 and 3.6; need to load a pre-defined clean and dirty data, and fixed set of denial constraints, making it not applicable for this research.
Raha and Baran	Source code available Not focused on Big Data and need golden standard of a dataset.	Provided a comparison method to compare KATARA, ActiveClean also and proved that none of these tools are adequate with Big Data
DISC	Used distance calculations and bounds to correct outliers into inliers to improve future clustering applications.	Provided a comparison with HoloClean and some other tools.

6.4 Evaluation and conclusion

Given that this research study is handling real world datasets with no knowledge of the ground truth for correct data values, the most adequate evaluation criteria is to use the plausibility criteria, similar to Chapter 5. On top of the plausibility metric, the proposed prototype was also benchmarked using the mean absolute error metric for the data imputation repair part as the

prototype made use of a linear regression algorithm. Other experiments applied source code obtained and allowed to benchmark some algorithms/tools such as KATARA, ActiveClean and Raha/Baran. As these algorithms were mostly classification tasks, the metrics used were precision, recall and f1 scores. In terms of accuracy of imputation scores, the prototype had a mean absolute error of 0.0 whereas Raha/Baran showed a precision score of 100%, while the other algorithms such as KATARA and ActiveClean denoted a much lower precision score at instance level of data. Thus, the prototype and Raha/Baran were much better in terms of accuracy, but in terms of time complexity, the prototype could perform imputation in a matter of few seconds on a local computing platform whereas Raha/Baran took almost 1 hour on a cloud platform. Thus, the data prototype was considered much better for data imputation repairs.

In another experiment, 'BayesWipe' could be replicated but showed that it did not execute on the whole first real-world dataset but could only perform data repairs in a reduced version of the dataset. Therefore, there was no actual need to perform the same experiments on the second and third datasets since it was already established that this tool did not perform well on Big Data. Another option discussed in existing data cleansing research studies is the possibility of using a crowd to perform the evaluation (Sushovan et al., 2014). However, for the purpose of this research study, this is practically unfeasible due to time and cost constraints.

Based on a combination of knowledge gained from literature surveyed (See Section 6.2) and experiments tried out (Section 6.3), some conclusions were derived to help optimize the level of data repairs for EHR Big Data. These conclusions inform the data improvement stage of the data quality methodological approach, by making sure the following steps or characteristics (listed below) are included.

1) Data cleansing tools or algorithms must be optimised to work with Big Data characteristics. Most of the tools or algorithms (discussed in Section 6.2 above) failed to perform data repairs on the first real-world dataset used, which contained more than 1 million rows of data and around 20 attributes. Most of those data repair tools or algorithms had not been designed to accommodate real-time streaming data sources. Cloud-based or parallel computing architecture may be needed to support proper data cleansing activities as the experiments upon Raha/Baran proved.
2) There is a need to ensure proper evaluation of corrected or transformed values of different types of data. From the constrained experiment with 'BayesWipe', the number of errors

generated by the cleansing process on a very limited dataset is alarming. Even if only one existing previous literature on the web did mention that 'BayesWipe' did not work properly with numerical values, there was no mention of this extreme limitation in the official documentations. Even with the experiments involving ActiveClean, KATARA and Raha/Baran, some of the metrics did not seem realistic and confirmed that with the lack of truth samples of data, metrics such as precision, recall and f1 score might not be the most adequate.

3) The process of data cleansing should be kept as simple as possible. With Big Data, it is not practical to generate a clean dataset/ data model separately from a dirty dataset/ data model. Hence, the use of artificial intelligence, or self-learning systems, should be promoted. This research study laid emphasis on and investigated how far machine learning could be helpful in this precise context. Some progressive learning of models might have been very useful, such as the one discussed with 'ActiveClean' software, but the results of experiments involving ActiveClean shows low level of accuracy of data imputation repairs.
4) Most of the algorithms/tools analysed in this chapter do not focus upon completeness and accuracy DQDs but made use of pre-defined rules or set of denial constraints. These violation detection data repair methods are not generally advised in Big Data context but may be a solution when the data model for the Big Data use case is known and stable.
5) As part of the data assessment and improvement stages of data quality methodological approaches, it would be useful if some exploratory analysis could be performed as part of data auditing. This would help in understanding data models, but also might unearth certain systematic types of errors. With this knowledge, machine learning algorithms might be optimised to deal with the specific DQ issues discovered with the exploratory analysis. In a more general case for this research study, it was found that catering for the completeness DQD is much more common than the accuracy DQD as discussed in Chapter 5. As proposed by both chapter 5 and overall literature relative to chapter 6, a degree of semi-automated or at least a minimum use of human expert intervention to vet accuracy errors detected and data repairs needed is necessary.

All the steps/activities/characteristics discussed above will inform the development of a proposed data quality methodological approach towards optimisation of data quality for EHR Big Data. The new approach will be compared with existing frameworks and methodological approaches using well-defined criteria in the next chapter.

Chapter 7: Proposing a methodological approach for optimising data quality in EHR Big Data

7.1 Introduction

This chapter discusses data quality methodologies/frameworks and recommends a proposed data quality methodological approach to optimise data quality in Big Data in the health industry. This proposed methodological approach is the main deliverable of this thesis and answers the research question "*What could be an optimized methodological approach to enhance DQ in Big Data for EHR data?*". The proposed approach aims to support practitioners in the field of Big Data quality (with special focus on EHR) to put into place *data driven data quality* also termed as *a posteriori* initiative (Srivastava et al., 2019). The proposed approach is based upon the knowledge derived from the results of the previous chapters of this research study combined with important features discussed as part of literature review in this chapter. In the initial sections, there is a comparison of existing data quality methodologies. Then, a detailed description of the proposed approach is made and finally, an evaluation of the proposed data quality methodology is discussed with an applied research approach. This applied research approach takes the form of qualitative research methods described in Chapter 4 and experiments described in Chapters 5 and 6.

Data is usually the product of a data strategy in organisational settings. In the best-case scenario, there is an alignment of the data strategy itself with an overall information systems (IS) strategy. The IS strategy is supposed to fit into an overarching business strategy. The operations of data in an organisation follows a life cycle and the main steps of data lifecycles are planning, obtaining, storing and sharing, maintaining, applying, and disposing of data (POSMAD) (Cichy & Rass, 2019). Hence, as data quality issues may be the result of activities in each of these above-mentioned lifecycle steps, there are many research studies that rightfully argue that data quality methodologies should be process driven (Batini et al., 2009). This effectively means that data quality is a holistic organisational endeavour, requiring human, technical and procedural inputs. However, process driven strategies towards data quality could be time consuming, costly, and tedious.

Another viewpoint of data quality strategy is **data driven**, where the aim is mainly to focus upon technical possibilities of improving the level of data quality wherever data is stored, either in databases, datasets or any other type of data repository. This effectively results in initiatives

aimed at improving DQ in datasets based upon knowledge of the data, metadata available, data exploration activities and data schemas. This type of strategy is less tedious, time consuming and costly. It can lead to datasets possessing a level of DQ which is thought to be good enough to be used by further software applications and data analytics tools and provide trustworthy outputs. These ideas mentioned form an important component of a new field of research known as *analysis driven data quality* in the data science domain (Glowalla et al., 2014; Baldassare et al., 2018).

Some of the techniques involved with data driven strategies are *acquisition of new data, data standardization, record linkage, data and schema integration, source trustworthiness, error localization and correction* and *cost optimization* (Batini et al., 2009***). This current research study followed a data driven strategy, and more particularly focused on cost optimization through a proper identification of DQDs involved and upon error localization and correction with the support of ML algorithms***. However, it is generally accepted that a combination of both process and data driven strategies certainly increases the level of DQ in any context. This is most probably why many traditional DQ methodologies and frameworks use a combination of both strategies.

7.2 Discussion of existing DQ methodologies

There are several existing research studies whose purposes are to discuss, survey and compare DQ frameworks and methodological approaches (Batini et al., 2009; Cichy & Rass, 2019). The differentiation between frameworks and methodological approaches is not very rigorous in those research studies; hence, the present research study also considers some studies that have been mentioned as frameworks. Even if their focus is not specifically upon Big Data, some of the existing research studies do extensively discuss either structured, semi-structured and/or unstructured data. Cichy and Rass (2019) even excluded some approaches that are industry specific from their survey, such as the CIHI, which was considered only applicable to the health industry and could therefore not be considered as generic DQ approach.

The criteria for the DQ methodological approach selection used for this thesis is comparable to what was done in the aforementioned two studies. The main components used by Cichy and Rass (2019) for comparison of approaches were data quality definition, data quality assessment

and data quality improvement. The DQ definition includes the understanding of DQ characteristics in terms of DQDs. The DQ assessment and improvement components include the description of steps to understand the nature of data quality issues and to propose repairs and/or transformations. Cichy and Rass (2019) also acknowledge the issue of lack of standards surrounding the discussion of DQDs as part of existing knowledge described in literature, as discussed in Chapter 4 of this thesis. Proper knowledge of DQDs is considered as the starting point to subsequent future successful DQ assessment and improvement phases.

On the other hand, Batini et al. (2009) used 5 main criteria for comparison of approaches in terms of (1) the methodological phases and steps, (2) the strategies and techniques, (3) the data quality dimensions, (4) the types of data, and, finally, (5) the types of information systems. They also argue that generic data quality methodologies possess the following three main sequence of activities:

*State reconstruction***:** it is an optional phase whose main aim is to collect contextual information that might support the data quality activities.

*Assessment/measurement***:** measurements are taken for DQD values, and assessments are performed when those measurements are compared against reference values to diagnose levels of data quality.

*Improvement***:** this concerns steps, strategies and techniques for reaching new and better levels of data quality.

As the current research study is based on Big Data, only those approaches that apply to structured, semi-structured and unstructured data combined were considered for further comparison in this research study. Hence, from the research study of Cichy and Rass (2019), the applicable approaches are AIMQ, COLDQ, TDQM and TIQM (complete names provided in Table 7.1 below). This ensures compliance with the variety characteristic, but unfortunately, there is not enough precision about the volume and velocity characteristics of Big Data. One of the results from Cichy and Rass (2019) is a decision tree acting as a guide to choose a possible approach. The decision tree asks the following questions:

- *What is the structure of data involved?* The idea is to choose only approaches that would support relevant types of data; in the case of Big Data, the conclusion would be choosing those that might support structured, semi-structured and unstructured data.

- *Which DQDs are relevant?* The aim is to use approaches that support targeted DQDs; in the present research study, the focus would be upon accuracy and completeness DQDs as objectively shown through the results of Chapter 4.
- *What type of measurements are preferred?* For the present research study, approaches supporting both objective measures, in terms of clear metrics, and subjective measures, in terms of opinions, would be preferred.
- *To what extent should costs be considered?* There are different types of costs that might be associated with data quality activities such as financial costs and performance costs. However, for this research study, financial cost will not be the most determining factor for approach selection.

The results of Batini et al. (2009) have broader applications as they provide some classification of approaches into four main categories but did not provide any precise idea for the selection of a potential approach. The categories mentioned are:

- *Complete:* these approaches provide support for the assessment and improvement phases, and address both technical and economic issues. The complete approach type could fit for this present research study due to it being heavily involved upon investigation of ML for DQ activities, however the emphasis of the current research study upon cost is negligible whereas it is an important element for the complete approach.
- *Audit:* these approaches focus more on assessment activities. This category is not adequate for the current research study since the improvement phase is highly important.
- *Operational:* these approaches focus on the technical sides of both assessment and improvement, but do not address economic issues. This category appears to be a perfect fit for the present research study.
- *Economic*: these approaches focus on the financial cost element, and therefore are not appropriate for the current research study.

The following table lists the selected approaches identified from existing research studies; the given acronyms will henceforth be used in this research study. These approaches are expanded upon based upon the current author's analysis of literature.

Table 7.1: List of DQ approaches considered

Acronym	Complete name of methodology	Main reference
AIMQ	A methodology for information quality assessment	Lee et al, 2002
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality)	Loshin, 2004
TDQM	Total Data Quality Management	Wang, 1998
TIQM	Total Information Quality Management	English, 1999
BDQPF	Big data Quality Pre-Processing Framework	Taleb et al, 2015
DQF4CT	Data quality issues in classification tasks	Corrales et al, 2018

The next step is to compare the above selected DQ frameworks/approaches/methodologies in the context of data driven strategies that meet the following conditions:1) They clearly consider all types of data to reflect better the variety characteristic of Big Data; (2) They can be considered as 'operational' according to the categories listed above by Batini et al. (2009), such that the focus is on both assessment and improvement steps and (3) they focus on some well-defined DQDs of accuracy and completeness.

AIMQ

One of the interesting points to note about AIMQ is that the focus is on 'information quality (IQ)' and not on data quality. It is made up of three major components: the PSP/IQ model, the IQA instrument and Gap analysis techniques. Very important IQ dimensions are grouped into four quadrants by the PSP/IQ model to facilitate decisions for improving IQ dimensions. The IQA instrument measures IQ for each IQ dimension. The Gap analysis technique assess IQ being used by an organisation for each of the four quadrants denoted by the PSP/IQ model (Lee et al., 2002). It is typically a process driven information quality strategy. The main aims were to assess information quality dimensions and perform gap analysis of those dimensions against organisations which possess optimal information quality ratings. Therefore, this approach does not really cater for improvements, but more for assessments. Using our comparison method and criteria in Table 7.2 below, this approach does not seem very adequate for Big Data quality for the health industry.

COLDQ

The goal of this methodology is to provide a DQ scorecard for the evaluation of the cost effects of low data quality. Improvement techniques, such as aggregation of costs during phase 6 of the

COLDQ methodology, generate direct benefits to organisations. Thus, this methodology fits more into the economic category of Batini et al. (2009), and therefore, does not seem very adequate for the data driven context of optimisation of data quality for Big Data in the health industry.

TDQM

This approach is based on the Deming cycle of 'Plan, do, check and act' and focuses upon information quality. It begins with a 'definition' stage that mainly identifies quality dimensions and IQ requirements. IQ metrics are produced during the 'measurement' stage. 'Analysis' stage identifies causes for IQ problems and assesses impacts of poor IQ. Finally, the 'improvement' stage proposes techniques to improve IQ. The comparison with our predefined conditions suggests that TDQM might not be totally adequate for Big Data, due to lack of clarity whether all types of data are supported and the fact that it does not seem to be a wholly operational category in terms of approach. It appears to be more of a *complete* approach, with the strategy being more process driven with less emphasis placed on the sole and independent data driven activities and steps, which have been explained earlier to be extremely important for Big Data.

TIQM

The focus of this approach is once more on information quality, rather than data quality. But, as the activities and steps are quite common across both data and information quality, this approach can be investigated further. However, this also implies that the quality strategy is more process driven, even if there are some data driven steps related with technical activities surrounding data improvements solely. It is made up of 6 main phases; phases 1 to 3 are more about understanding quality issues, assessing information quality and cost measurements. Phases 4 and 5 are about improvements, involving process improvements and correcting data. Phase 6 is overarching and is related with ensuring a proper information quality environment. As this approach was proposed late 1990s, it can be deducted that the focus would have been on structured data, since the use of unstructured data was extremely rare at that time. The overall conclusion from the evaluation of the criteria as listed in Table 7.2 suggests that TIQM is not adequate for Big Data.

BDQPF

BDQPF is referred to as Big Data pre-processing *framework* to cater for quality issues possible when attempting to apply data quality concepts to large datasets (Taleb et al., 2015), even if it is still a work in progress as many of the modules described had not yet been implemented and

tested at the time of writing of the study. It also considers data cleansing as the main Big Data pre-processing activity that directly affects data quality. Data cleansing refers to the process of searching, identifying, and correcting errors. Since this current research study also focuses on data cleansing, this makes the results from BDQPF very relevant. Some of the discussed methods involved with data cleansing are statistical, clustering, pattern based and outlier identification. BDQPF aims to be a framework which combines different components to increase data quality for Big Data, such as data provenance and data cleansing. The key components of BDQPF consist of the data quality profile selection, adaptation, and data quality control and monitoring.

BDQPF caters for all three types of data (structured, semi-structured and unstructured data) and focuses on the accuracy, completeness and timeliness DQDs. It is a very broad and extensive framework and discusses both process-driven and data driven steps and activities. Hence, it can be considered as an operational framework also. In terms of assessment steps, different types of rules such as 'auto-discovery', 'data domain' and 'user defined' or a combination of the three might be used to comprehensively detect and express the nature of data quality issues. Improvement steps are proposed to be DQD specific, such that there is optimisation of cost and data quality management. To conclude, this framework seems to be very promising to ensure proper DQ activities for Big Data, in the light of its comparison against our selected conditions. However, as this is a very broad framework, there needs to be much more precise and constrained steps proposed as part of an optimized data quality methodological approach.

DQF4CT

This is a framework to address data quality issues in data classification tasks. It is made up of: (i) a conceptual framework to provide the user guidance on how to deal with data problems in data classification tasks; and (ii) an ontology that represents the knowledge in data cleansing and suggests the proper data cleansing approaches (Corrales et al., 2018). The obvious difference between this framework and the current research study is that the latter aims to improve data quality levels for Big Data irrespective of its subsequent use. The main data quality issues associated with classification tasks were reported to be redundancy, timeliness, high dimensionality, duplicate instances, outliers, missing values and noise. Note that outliers and missing values are common with the current research study. Those two issues were subsequently categorized as noise with DQF4CT.

There are many data cleansing tasks proposed in this framework, but the focus of this section is limited to those tasks related with missing values and outliers. For missing values, the framework used either linear or Bayesian regression algorithms. **This aligns with the results from Chapter 5 of this current research**. For outlier detection, Local Outlier Factor (LOF) is recommended to detect the potential outliers. As discussed in Chapter 5, the current research study also proposed statistical based outlier detection methods for numerical values. However, DQF4CT does not mention anything about text value outlier detection. Conversely, one of the aims of this framework is to support data quality issues detection by non-experts, thus follows a somewhat semi-automated principle to data cleansing.

A diagram depicting the main steps of DQF4CT has already been provided as part of the literature review of this thesis (see Chapter 2, Section 2.2.11). Comparison with the three pre-defined conditions of this current research study concludes that even if DQF4CT confirms the need for proper data cleansing steps, especially relative to missing values and outliers, there are not enough details about whether the framework could be applicable for Big Data as there is no discussions on the types of data involved with the framework. However, important knowledge can be derived from this framework to inform the proposed methodological approach.

Table 7.2 below matches the approaches in Table 7.1 with the set of pre-defined conditions.

Table 7.2: Assessment of approaches against criteria for Big Data

Approach	Supports all types of data	Operational focus	DQDs considered	Assessment and Improvement steps
AIMQ	No, very unclear; none specifically defined.	No, seems rather 'audit' category	Yes, most of typical ones.	No; steps not really discussed.
COLDQ	No indication	No, is more a methodology focusing on benefits of improving data	Yes	Yes; more process driven approaches
TDQM	Unclear	Partially	Yes	Yes
TIQM	Unclear	Partially	Yes	Yes
BDQPF	Yes	Yes	Yes	Yes
DQF4CT	Unclear	Yes	Mentioned in terms of DQ issues.	Yes

From Table 7.2 above, BDQPF emerges as the most appropriate approach applicable in the Big Data context, and thus, will heavily influence a proposed methodological approach. One of the issues with BDQPF is that the steps discussed are too broad, whereas for the current research study the proposed data quality methodological approach aims to **provide more precise steps and activities**. The availability of these well-defined and precise steps and activities will allow DQ practitioners in the area of EHR Big Data to put into place clear and effective DQ operations.

7.3 Proposed data quality methodological Approach

Based on the discussions in the previous section and chapters, this research study proposes that an optimized data quality methodological approach for EHR Big Data should focus on the following three main components:

Prioritize important DQDs: any approach should develop clear and systematic techniques for understanding which DQDs are most important in a given use case or context. Most data quality approaches discussed above clearly lay the emphasis upon understanding DQDs or data quality issues as the foundational stage for any data quality initiative. Many of those approaches rely upon user and/or expert opinion for understanding DQDs. This could be applicable in applied contexts, such as for specific organisations or users. But for the development of more generic data quality solutions, human users might not be available nor reliable if ever available. In that case, the use of inner hermeneutic cycle (IHC) and Latent Semantic Analysis (LSA) are considered as cornerstones activities for this component. Although this research study focused on EHR, the researcher believes that the same techniques for determining the most important DQDs used in this current work may be applicable for other industries and use cases.

Detection of dirty data: this stage involves successfully and efficiently identifying dirty data, based upon the most important DQDs. As this current research study is focusing on data driven strategies, factors such as which information systems would make use of the data or requirements from final end-users are not considered in the process of detecting dirty data. To support the detection process, automated tools such as statistical calculations and machine learning algorithms are proposed. ML tools and implementation techniques must be carefully selected for Big Data, with the use of tools supporting out of core learning recommended. The support of ML depends heavily on implementation tools used, as the latter might provide practical constraints for ML deployment. Unsupervised learning algorithms are considered important, but supervised learning ones are not ruled out. However, to be able to deal with different types of accuracy

errors, a completely automated stage is deemed as inappropriate. There needs to be expert human/user involvement in determining whether outliers are errors. A final activity involves evaluating the detection process to ensure proper detection of dirty and correct data. For that purpose, plausibility measures are proposed due to the lack of knowledge about correct values highly probable with Big Data's velocity and variety characteristics.

Repairing of detected dirty data: this stage involves determining activities and techniques most efficient to correct dirty data. As discussed in the previous section, an optimized data quality methodological approach should be able to cater for all types of data. Most actual approaches fail, except for BDQPF on this aspect. Therefore, one of the activities as part of the improvement stage is to allow the ingestion of all types of data. Not all data repairs can be undertaken automatically specially in the context of Big Data but the use of machine learning coupled with a minimum amount of human intervention might be necessary. Models might be derived from ML algorithms, and this association with human expert knowledge might build more fault tolerant rules for data repairs. This is quite like the steps put forward by BDQPF. A final step is to propose proper evaluation of data transformations performed to ensure that no new errors are introduced as part of the raw data.

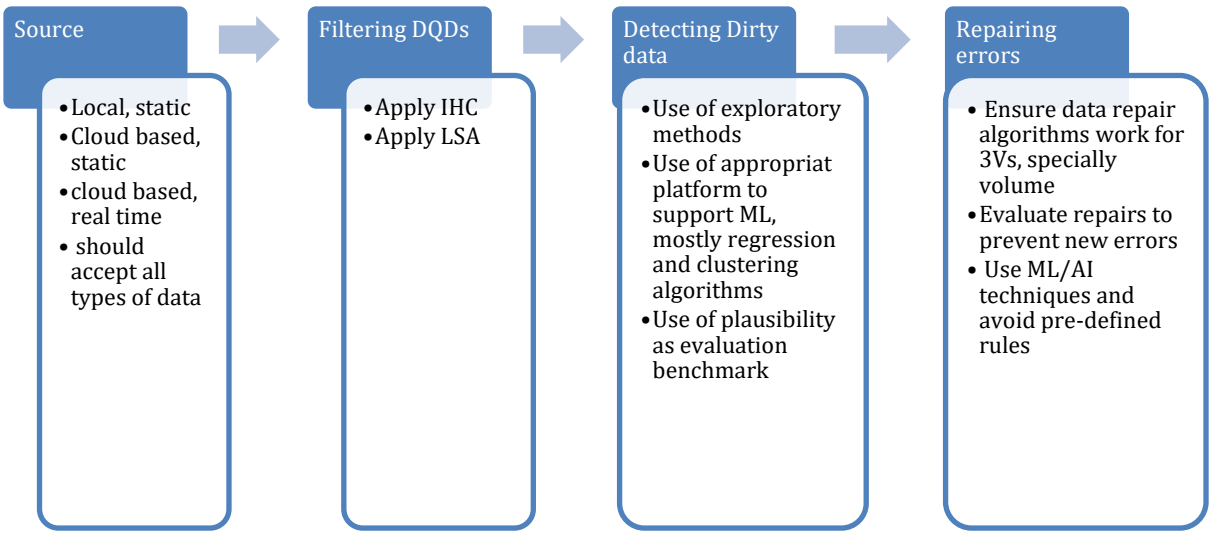

Figure 7.1: Main steps of proposed methodological approach (BDQMA)

The above diagram embodies the proposal of a methodological approach for improving data quality for EHR Big Data, denoted as BDQMA (Big Data Quality Methodological Approach). However, the steps are quite generic and can be applied to other contexts or industries, with potentially different results. For example, if different DQDs are discovered to be most important,

this could suggest support of other ML algorithms for detection of dirty data. On the other hand, completeness and accuracy DQDs are considered very important DQDs in general, and hence, it could be assumed that even if those results are focused towards the health industry, they might also be replicable for many other contexts.

The first component of the BDQMA is concerned with the source upon which DQ initiatives would be applied. During the investigations carried out as part of the literature review (Chapter 2 of this thesis), it is understood that Big Data sources might be either batch based, typically as part of a data lake, or real-time, for example as part of data streams coming from sensors. The learnings from Chapters 5 and 6 denote the impact that the technical architecture might have upon Data Quality initiatives. Hence, as part of the BDQMA, it is recommended to have DQ systems which would be capable of connecting with both locally based or cloud-based data repositories, and to be able to cater for both batch and real time data. The investigations connected with chapters 5, 6 and 7 specially show that most DQ algorithms, tools and approaches tend to specialize themselves upon one mode of data source only and hence will not be adequate in a Big Data context where all modes of data might actually be used. In the same logic, the first component of the BDQMA advocates that DQ systems should be able to cater for structured, semi-structured and unstructured data. This is due to the fact that as part of the investigations throughout this thesis, it had been found that many DQ algorithms, tools and approaches were limited to one type of data, and many limited only to numerical data within structured data. Such a recommended DQ system is a major technical challenge and might be very costly, financially and computationally, to put into place. Thus, the need to have BDQMA operate under a clear and well-defined organisational data strategy.

Second stage of BDQMA highlights the facts that a DQ system operating upon Big Data should prioritize certain DQD upon which high level DQ should be obtained to ultimately create a system with general level of high quality of data. A DQD provides a benchmark of what is meant by high quality data. E.g, completeness DQD means that the proportion of available and usable data must be very high relative to total amount of possible data. Thus, factors which might negatively impact the completeness DQD must be catered by a DQ system, and these factors might be broader and more numerous compared to causes of data errors. From Chapter 4, it was clear that the way to know about most important DQD was highly subjective. There was also a lack of uncertainty to what extent known DQDs could be relevant for Big Data systems in the healthcare area. Thus, the recommendation to apply IHC and LSA upon secondary sources to

confirm most important DQDs in a certain context of use. For EHR Big Data, accuracy and completeness DQDs were identified as most important, but this might be different in another context of data use, so ascertaining with IHC and LSA to focus upon specific DQD allows DQ systems to be very efficient.

The third component of the BDQMA is more constrained to the accuracy and completeness DQDs, with a focus upon missing data and outliers as potential DQ issues. To detect instances of these DQ issues, it is imperative to have a proper insight of data sources and hence proper and regular exploratory analysis must be carried out. Results from chapter 5 confirm Linear Regression as most effective for dealing with missing data, both in terms of quality of data imputation and time complexity. As a comparison, BDQPF advocates detection of DQ issues via the use of user-defined rules which have already been explained as not being realistic in a dynamic context such as Big Data. To deal with text data outliers, ML based clustering algorithms might support a human expert to detect the potential outliers more quickly. This component proposed plausibility as evaluation benchmark to use for DQ systems for EHR Big Data as the lack of golden standard of data prevents correct use of other regression and classification evaluation metrics.

The final component proposes steps to perform data repairs, which is an area that has been found to be given less attention compared to DQ detection. Results from chapter 6 denote that the few data repair algorithms cannot adequately carry out repairs on a high volume of data. The methods used by actual data repair algorithms depend upon pre-existing knowledge of how to correct dirty data through techniques such as Denial Constraints or Functional Dependencies. This knowledge is expected to be not available correctly with Big Data and therefore data repairs algorithms must have the capacity to learn from the data themselves. Hence, AI and/or ML based algorithms are recommended even if the results of chapter 6 points that use of AI/ML face inadequate computational complexity. Overall, this process of data repair is expected to be extremely challenging and custom-made AI/ML algorithms might provide better results. It is also recommended to properly evaluate proposed repairs by algorithms before accepting them as many of the algorithms/tools investigated as part of chapter 6 did not convince with the quality of the proposed repairs, either with Raha/Baran or BayesWipe experiments.

7.4 Evaluation of the optimized methodological approach (BDQMA)

The BDQMA is evaluated in this thesis via an experimental research method, where the main components proposed as part of this methodological approach were investigated and evaluated. The results of the evaluation have been discussed in detail within Chapters 4, 5 and 6 of this research study, but this section would highlight the key results and provide a critical reflective insight of each step.

The first step of the BDQMA is about connecting with a proper Big Data source. One of the main initial challenges of the current research study was determining how to appropriately reflect Big Data, through its main characteristics of Volume, Variety and Velocity. Most research studies in the field use CSV based datasets, but fewer and more recent studies started to use cloud-based datasets running on platforms such as Hadoop. The size of data storage also diverges in available studies, but for the experiments carried out in this current research, three large CSV based datasets and 1 BigQuery public dataset were used. The recommendation for the BDQMA is that any data quality solution should be able to connect with either a cloud-based or local source of data. Ideally, the choice and/or combination of both should be allowed. Furthermore, the data source should not restrict any type of data, but should be flexible enough to accommodate structured, semi-structured and unstructured data. The benchmark datasets chosen reflected real world data in a specific health industry use case, and even if there is a mixture of numerical and categorical data, it consisted of largely structured data.

The second step of the BDQMA is related to the assessment stage of DQ approaches, where the main goal is to understand precise data quality issues in a specific context. The investigations part of Chapters 4 and 7 of this current research study have made it clear that this precision starts with knowing which DQD to deal with. Thus, the BDQMA proposes that knowledge of most important DQDs for Big Data in any context be carried out by a combination of the use of Inner Hermeneutic Cycle (IHC) and Latent Semantic Analysis (LSA) techniques for developing general solutions. In the case that DQ solutions are being tailored for specific organisations or use cases, then involvement of users to understand the most important DQDs might also be envisaged. However, as a methodological approach, the BDQMA takes a generic perspective. Chapter 4 has detailed the experiments carried out with the use of both IHC and LSA and has shown that these techniques help pinpoint most important DQDs in a systematic, objective and non-bias way.

The third step is of the BDQMA concerns detecting data quality issues as part of a given data source and is based upon DQDs being considered. The experiments carried out as part of Chapter 5 of this research study has demonstrated that this detection process benefits from the use of ML algorithms but might also make use of direct statistical calculations and use of human judgement in some situations. Regression algorithms were found to be most effective for data imputations of missing values to cater for completeness DQD, similar to results of other studies in the data imputation domain as detailed in chapter 5. Clustering algorithms were determined more effective for detecting data outliers for text-based data as part of accuracy DQD. Throughout the general investigation on DQDs as part of Chapters 2 and 4, accuracy and completeness are very common across different industries and contexts, and hence, the use of the above-mentioned ML algorithms can be relevant in many cases. Also, with Big Data there might not be 'truth' samples for evaluating ML algorithms. Therefore, well-known ML evaluation benchmarks based upon positive and negatives rates such as precision and recall are not realistic, and plausibility is a better benchmark for evaluation.

The fourth step, concerned with repairing detected errors, was evaluated in Chapter 6 of this thesis. A prototype, including the techniques learnt from Chapter 5, was compared against some data cleansing algorithms/tools discussed in literature.. The comparisons denoted the almost unusable time complexity of certain data repair algorithms and the efficiency of linear regression for data imputation. The rationale for giving the onus upon human judgement for final data accuracy repairs, both for numerical and text data, was also explained and supported. However, the knowledge gained from those experiments and discussions from literature resulted in the ideas forming this fourth step, which are 1) ensuring that Big Data characteristics, especially the volume property, are taken into account for data repairs solutions, which does not seem to be the case currently, 2) the need to properly evaluate corrections made, such as to ensure new errors are not being introduced and 3) as far as possible, apply automated data repairing techniques based on ML or other forms of AI, but in any case, avoid making use of pre-defined rules since the nature of errors cannot be exactly pre-determined with Big Data.

7.5 Conclusion

This chapter investigated existing data quality frameworks, approaches and methodologies with the principal goal of proposing a new optimized data quality methodological approach adequate for EHR Big Data. It was named BDQMA (Big Data Quality Methodological Approach). Even if empirical experiments as part of evaluation for this research study was set for EHR, one key decision factor behind proposing the steps of the data quality methodological approach was whether a particular step could also be applied in other contexts or industries. The strategy pushing this current study's proposal is data driven. The choice of this strategy also heavily impacts on steps which are proposed as part of the methodological approach.

The review of literature in the field of data quality denotes that there are many frameworks, approaches and methods discussed over the years. Most of those are not suitable for Big Data, since they focus mostly on structured data and follow process driven strategies. However, it is only recently that research focused upon Big Data have emerged, such as the BDQPF framework. **One of the main contributions of this research study is the fact that the proposed BDQMA is a much more well-defined and precise methodological approach compared to BDQPF, consisting of clear steps and activities, with the aim of optimizing data quality for Big Data.** A comparison of the existing approaches was performed using four main factors (Section 7.2 above), which have been inspired by well-established research in the field of data quality in general but catering for Big Data compliance.

Comparing BDQMA with BDQPF might adequately highlight the added value which BDQMA brings to knowledge in the field as BDQPF has already been assessed as being the most appropriate approach from the analysis of approaches in table 7.2. The first main stage in both BDQMA and BDQPF is concerned with data sources, and they are quite comparable as both advocate for the data sources dealing with different structures of data. Even if BDQPF does not explicitly state it, it can be assumed that BDQPF recommends being able to connect to both local and cloud-based repositories also. The second stage of BDQPF is known as DQ class selection and is made up of 12 sub activities. It discusses about the need to perform activities such as data cleansing, data integration and data enrichment but no exact precise details about how to perform them are provided. Hence, many activities are advocated within this stage but not precisely explained how they should be carried out under different contexts. This is very different from BDQMA which recommends only two activities for its second stage, and which have been

adequately detailed. It can also. BDQPF carries out its actual data pre-processing activities through the use of different rules (user defined, domain related and auto discovery). Through chapters 5 and 6, it had been made clear that the use of rules should have adverse negative effects upon the performance of DQ systems. Conversely, BDQMA proposes the use of a combination of statistical methods, ML/AI algorithms together with some human validation for its dirty data detection and data repair stages. This combination should result in DQ systems having higher performances. In its approach, BDQPF is more process driven whereas BDQMA is data driven, thus BDQMA is more well-defined.

The BDQMA was evaluated throughout this thesis especially in Chapters 4 to 6. Knowledge gained from those evaluations/ experiments validated steps and activities forming the BDQMA, with critical discussions upon their adequacy for Big Data. Overall, the BDQMA could be used as a guide for stakeholders wishing to carry out a data driven strategy for optimizing Big Data quality. It opens the door for further evaluations which might be implemented in different industries and contexts as part of potential future research.

Chapter 8: Summary & Conclusions

This chapter will present the main outcomes of this research study in a more concise and succinct format to understand the importance of the results, the novel contributions to knowledge, and explain reflective comments and recommendations. It is well accepted that DQ is a very contextual domain, and as such, there cannot be generic approaches, frameworks, activities and steps which could be common in all industries and all use cases. This led the research study to focus on one industry, in terms of the health industry and its possible data use. This chapter addresses how far the research question and objectives have been met. To recap, the research question set out was:

What could be an optimized methodological approach to cater for Data Quality for EHR Big Data?

The three main research objectives part of the framework guiding the methodological approach development were:

a) *To determine what are the most important DQDs of EHR Big Data*
b) *To investigate the appropriateness of machine learning algorithms for detection of DQ issues for EHR Big Data*
c) *To investigate appropriateness of data repair algorithms for EHR Big Data*

8. 1 Summary of results and novel contributions

This section is discussed as per the research objectives (RO) listed above, ultimately reflecting upon the research question.

RO1: To determine what are the most important DQDs of EHR Big Data.

The rationale guiding this research objective lies in the proliferation of DQDs being discussed as part of current knowledge, and the lack of precise knowledge surrounding the importance of DQDs in specific contexts. Understanding which DQDs are most important is the foundation of any proper DQ approach, as evidenced by literature discussed in Chapter 7. The lack of knowledge of most the important DQDs in a specific context necessitated systematic and unbiased research methods for secondary data, in the form of IHC and LSA. The IHC is a more qualitative based method, ranking the importance of DQDs via a systematic and integrated literature review. It may, however, be affected by the researcher's bias, and therefore, a more

quantitative method, the LSA, was applied on the same corpus of literature identified through the IHC to triangulate the IHC results.

In terms of results, 46 different DQDs were highlighted through the IHC, with *accuracy, completeness, consistency, reliability and timeliness* found to be most important according to a weighted count method applied. However, with the LSA method, only 10 DQDs denoted significance in the same literature corpus, with *accuracy, usefulness, completeness, availability and validity* found to be most important. From those results, *accuracy* and *completeness were* the common most important DQDs, and were used as the basis for DQ issues detection and data repairs as part of experiments to inform a potential data quality methodological approach. Only these two DQDs were chosen in the context of this research study as they were significantly more important than the other DQDs identified, especially with the IHC method where accuracy showed a weighted count of 73 whilst completeness was 71. The next highest DQD's weighted count was 47.

Thus, a data quality methodological approach to identify most important DQDs in a given context can adopt the combined use of IHC and LSA. These methods work adequately on secondary data, and thus can be transposed to many different industries and use cases. It is ideal in contexts where there are no users or experts who can pinpoint the most important DQDs directly, which is more expected for tailor-made organisational DQ activities. But, whenever DQ experts or solution builders aim to create DQ solutions as off-the-shelf packages, then the two steps of IHC and LSA should be very effective in ascertaining the most important DQDs and guiding appropriate DQ actions.

Novel contributions: Ranking of DQDs in Big Data investigations is an emerging area of research. This section of the current research study proposed rigorous and systematic research methods to rank most important DQDs. Existing studies were based on more subjective methods and to the best of the current author's knowledge, there is no previous study which ranks the importance of DQDs. The results confirmed that well known DQDs, accuracy and completeness, are still very important for EHR Big Data.

RO2: To investigate appropriateness of machine learning algorithms for detection of DQ issues for EHR Big Data

Due to the characteristics of Big Data, traditional DQ issues detection methods such as the use of conditional functional dependencies (CFDs) are not relevant in Big Data contexts; firstly, due to multiple provenances of data, the exact nature of DQ issue might be unknown and secondly, applying those CFDs on the volume of data might lead to unusable response times. As ML is becoming more and more prominent, this research study aimed to explore how far ML can be useful to support DQ activities, particularly the use of ML algorithms to detect dirty data. It must be noted that the use of ML in view of supporting DQ is a nascent domain, and one novelty of this research study is that it is probing deeper in this new domain in the very specific context of EHR Big Data .

As stated earlier in this chapter in the discussions about RO1, the DQ issues related to EHR Big Data are represented by the accuracy and completeness DQDs. More concretely, completeness issues can be exemplified with missing data whereas accuracy issues can be exemplified by outliers. This research study made use of real-world data as part of experiments for assessing effectiveness of different ML algorithms and other statistical tools. The characteristics of the selected datasets used for the experiments are detailed in Chapter 5.

Relative to missing values detection, ML might not necessarily be required as many databases and statistical tools possess features to detect missing values. Therefore, this research study focused upon use of **ML to impute** missing values. ML use faces several challenges to cater for imputation, such as the volume aspect requiring some ML algorithms to execute properly only on a cloud based architecture. However, with the experiments which were possible both on local and cloud-based platforms, linear regression stood out to be more plausible for EHR Big Data, confirming current discussions about efficiency of Linear regression for data imputation activities within current existing literature. Another important finding is that the evaluation of ML imputation algorithms of real-world data cannot use typical benchmarks such as recall, precision and harmonic means, but plausibility is the most suitable evaluation benchmark. The use of ML to support imputation of missing values for Big Data is again a nascent research topic, with very limited previous studies.

Statistical algorithms, such as ‘z-score’, were found to be very efficient in detecting numerical outliers as part of activities to handle issues associated with accuracy DQD. Thus, use of ML is not advised for detection of numerical outliers. With text data, detection of outliers requires a

process of data transformation into ‘TF-IDF’ scores, and subsequently, clustering-based ML algorithms such as k-means, are adequate to help the detection of outliers. It must be noted that the issue with outliers is that they refer to data points which stand out from the bulk of other data points, and this research study concludes that the final decision of ascertaining whether an outlier is an accuracy error needs to be taken by a human user, especially whenever critical data use is involved. The support from ML is that it helps to pinpoint or identify those outliers efficiently.

As opposed to the recommendations to know the important DQDs, the detection of DQ issues recommended by the data quality methodological approach steps discussed above cannot be generalised to different industries and data use cases, due to the potential differences in types of data involved. This is in-line with the ‘impossibility theorem’ related with ML algorithms, which discuss about the fact that the same ML algorithm produces different results based on different datasets and parameters involved (Pandove et al., 2018). However, since numerical and text data were considered in this study, it can cautiously be assumed that these steps might prevail in many other contexts where these types of data are used. However, for Big Data use in other cases, further avenues of research might explore use of ML algorithms to other categories of DQ issues as part of other types of data.

Novel contributions: The in-depth discussions surrounding the potential application of AI/ML algorithms for detecting dirty data represent a novel contribution in the area. Challenges of use of AI/ML for dirty data detection firstly related to the technical architecture, whether related to the technology used to implement the algorithms or connected with data types or associated with the volume of data. The use of real-world datasets (both CSV and BigQuery) provided the experiments a greater empirical impact, together with the systematic selection of algorithms discussed with existing research literature. The results proved that ML algorithms must be combined with statistical tools and minimal user involvement to enhance the dirty data detection process. Practitioners and researchers around the use of ML for DQ with Big Data may refer to the steps as per BDQMA to guide their initiatives.

RO3: To investigate appropriateness of data repair algorithms for EHR Big Data

As with DQ issues detection, data repairs can be different and more challenging with Big Data due to the 3 V’s characteristics. One example is that some supervised learning-based methods,

such as ‘BoostClean’, are used for data repairs according to existing research studies (see Chapter 6, Section 6.2). However, it is generally considered that a labelling process is a real challenge with Big Data. Thus, as part of a data quality methodological approach for Big Data, it is important to understand and evaluate adequate and efficient ways to carry out data repairs, which form an essential component of the data improvement stage proposed by many existing DQ frameworks. Data repairs for Big Data are a novel domain of research, and therefore the amount of existing research studies in the area is relatively limited.

One of the first major conclusions from the experiments carried out is that there were few available tools and algorithms which could be used to repair data in the benchmark datasets.. Most of the existing tools and algorithms failed to process 1 million rows of data, but could process around 18 000 records of the second CSV based dataset.It is evident that the volume of Big Data is currently a major stumbling block for data repairs. Furthermore, most tools do not allow any type of real time streaming data connection, which hinders the velocity aspect of Big Data also.

In one experiment involving a pioneer data cleansing solution, BayesWipe, only a small scale of the benchmark dataset could be executed. Furthermore, it is noted that the data repairs resulted in a consequent number of errors. Hence, as part of a series of steps for data improvement, it is very important to crosscheck proposed data repairs before enforcing them. As noted in RO2, evaluation with real work data can itself be very challenging and constitutes a very specific area of research for the Big Data quality domain.

Many existing data repair algorithms and tools base themselves upon techniques such as DQ rules and/or denial constraints. These techniques are constraining and unrealistic with Big Data as it implies a fixed and stable pattern and distribution of data, which might not always be true for Big Data. Thus, even if ML-based data repairs might be considered as being a new area, there are some techniques such as the use of progressive learning of models and use of ensemble learning which might become adequate ML-based data repair solutions. One experiment involving progressive learning of models through the ‘ActiveClean’ algorithm also pointed to high time complexity and unrealistic performance results. This thesis argues for the exploration and use of potential ML based techniques for data repairs as better and newer techniques become available.

The data repair activity still requires the potential use of human judgement in terms of specifying the most appropriate data replacement values, specifically with outlier corrections. This will help in terms of minimising the number of new errors being introduced after data repairs. In general, automated ML powered data repair methods need to be considered but investigations in Chapter 6 denote that the proposed automated corrections might either need an external knowledge base for text data and replace with incorrect inliers for numerical data Therefore some degree of human involvement is still recommended to be critical specially in sensitive data contexts.

Novel contributions: Apart from also being an area with relatively few existing focused research, this section of the thesis highlighted and confirmed the difficulties by existing algorithms and tools to perform data repairs for Big Data with the volume aspect representing a major problem. It also highlighted various other avenues of research and exploration which needed to be carried out with more resources and focus. However, even if the use of ML seems promising for data repairs, complete automation of those repairs needs to be undertaken carefully. The use of user intervention, though impacting response time, may be in some cases necessary to validate data repairs done.

RQ: What could be an optimized methodological approach to cater for Data Quality forEHR Big Data?

This is the principal research question of this thesis, in terms of proposal of a methodological approach to optimize DQ for EHR Big Data. The BDQMA consisting of four main stages is proposed, explained and justified in Chapter 7 as the principal deliverable of this research study. Current knowledge denotes a lack of frameworks, approaches and methodologies related to data quality for Big Data. One recently proposed framework (BDQPF) which takes Big Data characteristics in mind is too broad and hence lacks the specificity needed to drive DQ initiatives with Big Data. Most of the different components forming the BDQMA were already evaluated through different methods detailed in Chapters 4 to 6. The components give precise steps and activities that must be followed to undertake proper DQ for EHR Big Data. Although the aim of those steps is meant to be as specific and precise as possible, they are technology neutral and independent, even if some experiments were carried out using certain tools and platforms. The focus of this research study is upon EHR data, however, the discussions in the explanation of BDQMA (Chapter 7) explains clearly which components might be generic and which ones are

specific to EHR context. As per the objectives of this current research study, taking into account the scope, goals and constraints, it is argued that the BDQMA is an approach which is data driven and provides more precise and concrete steps/activities to support a DQ pipeline for EHR Big Data .

Novel contributions: The inadequacy of most existing data quality frameworks for Big Data application following a data driven strategy has been discussed and explained. However, one recent proposed framework, BDQPF, had provided some encouraging results according to the evaluation criteria used in Chapter 7, but is more processed-driven and is quite abstract. On the other hand, BDQMA achieved the goal of providing very clear and specific steps and activities for its four main stages and is purely a data driven approach. Thus, practitioners of EHR Big Data may adopt BDQMA to increase their confidence of improving DQ levels irrespective of the future use of the data concerned.

8.2 Evaluation of work

This work followed a classical 'funnel' shaped research drive. That is, the research area was originally very broad, and got refined and focused to more specific research questions and objectives as knowledge of the different areas concerned increased. The initial goal was to perform research focusing on Big Data governance. A background study of the area concluded that Big Data governance is made up of too many broad steps, and many of those individual steps could each be a PhD focus. Thus, the decision was taken to refine the work to focus on a particular aspect of Big Data governance, in terms of Data Quality for Big Data. At the start of this research study, research on DQ for Big Data research was rare, and although there are now more research studies in the area, it can still be considered a growing and buoyant research area. DQ is a quite well-established domain, for traditional database platforms, however, since Big Data is a new technological initiative, there were large gaps in terms of research knowledge. Using the 'funnel' approach allowed the researcher to understand the broad topic of Big Data governance, technical aspects of Big Data (such as storage and analytics), data quality techniques used in traditional databases and machine learning/ deep learning algorithms. Even if those ideas are peripherally related to the exact research objectives, it was important to understand them to develop a comprehensive research question which fills some research current gaps in the area. One of the general challenges faced for a research study spanning many years is the fact that knowledge in the field gets updated quite rapidly, and there is always the need to frequently check for updates to incorporate them into the research study.

Furthermore, the researcher discovered that DQ is a largely contextual process, that is, DQ activities, methods, tools, and techniques can hardly be generalised in all contexts. This fact brought about constraints to the research study which subsequently specialized upon one area, in terms of EHR. While probing more into data being associated with the health industry (a recognised adopter of Big Data), the researcher also discovered that that the array of types of data involved was broad. Since the aim of the research study was to develop an optimized methodological approach, the decision was hence taken to focus upon real-world data, and therefore experiments were carried out using EHR datasets, both on local computing and cloud based architecture. Nowadays, replicating Big Data is more affordable with available tools, and this research implemented experiments involving cloud based BigQuery Datasets involving several data types..Those cloud based experiments were involved as part of chapters 5 and 6 involved with the data assessment and data improvent sections of typical DQ pipelines. Those experiments helped to confirm the proposed steps and activities recommended through BDQMA.

Getting access to primary source of use of EHR Big Data was one challenge as health organisations were very cautious about giving access to their data citing the sensitivity of the data and data protection laws. Furthermore, Therefore, this fact provided some challenges in terms of being able to carry out case studies and/or real-life experiments, and in terms of access to Big Data practitioners and experts. The choice of the research methods employed in the different sections of the research study was hence oriented towards secondary, literature-based methods for the investigation of DQDs and use of publicly available datasets, tools and algorithms, for the Big Data quality detection and repairs experiments. The use of secondary data had the benefit of allowing the work to be as generic as possible and not biased by under-sampled opinions from experts.

This research study investigated the potential of use of ML algorithms for data improvement stage, specifically for the activities of DQ detection and data repairs. By constraining the DQ issues to two most important DQDs, the aim was to reach a better understanding of the potential effectiveness of certain ML algorithms to optimize DQ activities in a specific sector. The literature survey performed on use of ML for detecting completeness and accuracy issues in EHR Big Data demonstrated the scarcity of existing studies. Hence, the search was made broader, either with the use of ML for normal data, or use of ML for other use cases different from health contexts. The use of ML algorithms as miracle solution was quickly discarded. Most discussed

representations of the two DQDs (accuracy and completeness), in terms of outliers and missing values in general, were investigated in terms of the suitability of ML algorithms to tackle them. Even if there were some proposals from existing research studies, some of the articles and conference papers consulted lacked the ML implementation details to replicate them. . Thus, this research study implemented some of the discussed ML algorithms in existing studies via different tools and programming languages, such as 'RapidMiner Studio', ,python (versions 2.7, 3.7 and 3.8), Vertex AI and BigQueryML amongst others. Even if some of the ML algorithms could not be replicated as in the original articles, attempts were made to come up with algorithms which were quite close and respected the general logic of the required algorithm. A lot of knowledge was obtained from the development and implementation of ML algorithms, especially on real-world data and in a Big Data context. It was very surprising to see how many of those tools and algorithms could not cope with the volume of the benchmark databases. The 'curse of dimensionality' is a well-known challenge for high dimensional data. E.g., this issue affects clustering algorithms such that distance-based measures are known to be not effective (Pandove et al., 2018). The current research study hence confirms this 'curse of dimensionality' issue with clustering algorithms involved in the experiments made. Even the method of evaluating effectiveness of some ML algorithms for detection of DQ issues had to be different from the widely cited 'precision and recall' benchmarks as there was a lack of 'ground truth samples' related to the known correct values of data items.. All the knowledge derived from the experiments informed the development of BDQMA.

The components of the methodological approach involved with data improvement cannot be generalised to different types of data. For example, detection of outliers for numerical data could quite easily be achieved through known statistical methods but detection of text data was more adequately supported by clustering-based algorithms together with human expert involvement. Even for data repairs, this research study found relatively few existing tools and algorithms which could account for data repairs for Big Data. One tool known as Raha/Baran performed some level of data repair but took hours to execute and was therefore considered as not very adequate for Big Data. In another experiment, the 'BayesWipe' algorithm failed due to the 'Banjo' algorithm not being able to accommodate the first benchmark test dataset. However, even if some algorithms/tools faced implementation issues, the failed and/or incomplete experiments resulted in gaining knowledge which informed the data quality methodological approach (in the same way as the successful experiments). Thus, those challenges did not hinder the proposal of the

methodological approach but have opened more focused and different avenues of future research possibilities in the research area.

Overall, the BDQMA aligns well with one of the few existing DQ frameworks for Big Data (i.e. BDQPF). The justifications for the different steps of the BDQMA are also correctly sustained, despite the challenges discussed. BDQMA follows typical DQ pipeline components, but is focused on data driven approach and provide clear and precise steps/activities to perform. This is different from what frameworks such as BDQPF advocates.

8.3 Limitations of work

The first limitation of the research is related to the context in terms of access to Big Data quality experts, both in terms of use and implementation, and organisations (more particularly connected with the health industry) that have adopted Big Data.. The ideal situation would have been to get into contact with experts and organisations who were working on the domain of data quality for Big Data, but this was practically unfeasible in terms of cost and researcher convenience. Some attempts were made to contact potential research experts in Big Data during conferences or through research platforms such as 'Researchgate', but the very few responses obtained pointed out that the DQ experts were not typically involved with the field data quality for EHR Big Data.

Another less important limitation is that the research was not funded, and therefore some tools which may have been part of some experiments were prohibitively expensive. The researcher therefore had to rely on freely available versions of certain software and tools. The use of GCP with $300 credits proved to be enough to carry out experiments relative to chapters 5 and 6 mostly.. This, however, did not impact the ability to propose the data quality methodological approach.

A final minor limitation is related to the use of freely available external real-world test datasets being used as part of the experiments. As explained in the precedent section, it was important to use real-world datasets to carry out experiments more precisely to determine the effectiveness of ML algorithms, data repair algorithms and software tools. Unfortunately, this resulted in a lack of 'ground truth' samples for the evaluation of the different algorithms. However, the benchmark used for evaluating the experiments (i.e., plausibility), is a well-accepted benchmark in such circumstances.

8.4 Recommendations for Future work

One area of future work might involve using different types of data sources as opposed to mostly numerical and text data as part of typical EHR datasets. Furthermore, instead of the static CSV-based and BigQuery real-world benchmark datasets used in the current research study, future research can explore data quality in data streams also. This will be possible only when real-world data sources will be more available for EHR data. The researcher or research team carrying out this type of experiment should also possess a consequential amount of funding. Another possible

reason for using new cloud based Big Data architectures is that future improved methods and techniques may be available to create synthetic datasets (used on these Big Data architectures). The use of synthetic datasets will allow the application of more classical evaluation benchmarks used in the area and might allow better comparison of the effectiveness of algorithms or tools.

This current research had the aim of comparing independent ML algorithms for DQ issues detection, focused upon missing data representing completeness DQD and outliers representing accuracy DQD. However, some recent research in terms of use of ML for detecting the above mentioned DQ issues tend to point towards the use of ensemble learning methods, which typically make use of a series of weak learners or ML algorithms. The use of deep learning algorithms, as supported with the experiments with GAIN, could also prove to be quite effective upon richer type of data made up of more unstructured data compared to the EHR datasets used for the experiments in the current research study. Therefore, a potential area of future research will be to include the use of ensemble learners in the experiments for both detecting DQ issues and for data repairs. Another area of future research related to the experiments might involve following an analysis-driven data cleaning approach rather than a data-driven data cleaning approach as done in the current research study. Analysis-driven data cleaning aims to improve data solely for the purpose of improving results of potential data analytics activities. Hence, the objectives of the data cleaning or repairs become different from those following a data driven approach.

A final area where future work could be undertaken might involve the use of practitioners and experts in the field of DQ for Big Data. These experts might be used to validate the results obtained using IHC and LSA to determine the most important DQDs in a given context. However, it would depend largely on the competency and ability of the experts to be able to formulate opinions relative to DQDs in a general context, for example, to reflect the DQ needs of a whole industry, rather than be constrained by the DQ needs of their own organisations.

References

Abdullah, N., Ismail, S.A., Sophiayati,S. & Sam M., 2015. Data Quality in Big Data: A review. *International Journal of Software Computing Applications*[online], Vol. 7, No. 3. ISSN 2074-8523

Aday,L.A & Cynamon M. , 2010. Health Survey Research Methods. *9th Conference on Health Survey Research Methods, Centers for Disease Control and Prevention* National Center for Health Statistics Hyattsville, Maryland

Abouzeid,A., Bajda-Pawlikowski, K., Abadi, D.J., Silberschatz, A. & Rasin, A., 2009. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *Proceedings of VLDB*, 2(1):922–933

Ahmed J. & Soomrani,R., 2016. TDTD: Thyroid Disease Type Diagnostics. *2016 International Conference on Intelligent Systems Engineering (ICISE)*, Islamabad, Pakistan, pp. 44-50, doi: 10.1109/INTELSE.2016.7475160.

Aisling O'Driscoll, J. D. R. D. S., 2013. 'Big data', Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, pp. 774-781. Volume 46, Issue 5.Elsevier. https://doi.org/10.1016/j.jbi.2013.07.001

Amoakoh-Coleman, M.; Kayode, G.A.; Brown-Davies, C.; Agyepong, I.A.; Grobbee, D.E.; Klipstein-Grobusch, K.; Ansah, E.K, 2013. Completeness and accuracy of data transfer of routine maternal health services data in thegreater Accra region. *BMC Res. Notes*, doi:10.1186/s13104-015-1058-3.

Amr, E. et al., 2013. NADEEF:A Generalized Data Cleaning System. *Proceedings of the VLDB Endowment*. Trento, Volume 6, Issue 12, https://doi.org/10.14778/2536274.2536280

Anderson, J.G., Ramanujam, R., Hensel, D., Anderson, M.M. & Sirio, C.A., 2006. The need for organizational change in patient safety initiatives. *International journal of medical informatics*. 75(12): 809-17, https://doi.org/10.1016/j.ijmedinf.2006.05.043

Arts,D.G.T.,De Keizer,N.F. & Scheffer,G., 2002. Defining and Improving Data Quality in Medical Registries, *Am Med Inform Assoc*.;9:600–611. DOI 10.1197/jamia.M1087

Baldassarre, M., Caballero, I., Caivano, D., Rivas, B. and Piattini, M. , 2018. From Big Data to Smart Data: A Data Quality Perspective. *Proceedings of the 1st ACM SIGSOFT International Workshop on Ensemble-Based Software Engineering (EnSEmble '18)*, November 4, 2018, Lake Buena Vista, FL, USA. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3281022.3281026

Batini, C., Rula, A., Scannapieco, M. & Viscusi, G., 2015. FROM DATA QUALITY TO BIG DATA QUALITY. *Journal of Database Management*, Volume 1, pp. 60-82.

Batini, C. & Scanapieca, M., 2006. Data Centric Systems and Applications. *Data Quality.* Springer.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., 2009. Methodologies for data quality assessment and improvement. *ACM Comput. Surv*. 41, 3, Article 16, 52 pages. DOI = 10.1145/1541880.1541883

BBC, 2014. *NHS fraud and error 'costing the UK £7bn a year'.* [Online] Available at: http://www.bbc.com/news/health-26654001 [Accessed 4 January 2016].

Becker, C., Duretec, K. and Rauber, A. (2017). *The challenge of test data quality in data*

processing. J. Data and Information Quality 8, 2, Article 7 (January 2017), 4 pages.
DOI: http://dx.doi.org/10.1145/3012004

Becker,D., King, T. & McMullen, B.,2015. Big data, big data quality problem, *2015 IEEE International Conference on Big Data (Big Data)*, Santa Clara, CA, USA, pp. 2644-2653, doi: 10.1109/BigData.2015.7364064.

Bertossi, L. & Milani, M., 2018. Ontological Multidimensional Data Models and Contextual Data Quality. *J. Data and Information Quality* 9, 3, Article 14 (March 2018), 36 pages.
DOI:https://doi.org/10.1145/3148239

Beyea, S.C. & Nicoll, L., 1998. Writing an integrative review. *AORN Journal*, Vol 67, Issue 4, https://doi.org/10.1016/S0001-2092(06)62653-7

Blake, R. & Mangiameli, P., 2011. The effects and interactions of Data Quality and Problem Complexity on Classification. *ACM Journal of Data and Information Quality*, 2(2). https://doi.org/10.1145

Bollier,D., 2010. *Perils and promises of Big Data*, Washington, DC. The Aspen Institute.

Borthakur, D., Gray,J & Sarma, J.S., 2011. Apache Hadoop goes real time at Facebook. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, SIGMOD 2011, Athens, Greece, June 12-16, 2011. DOI: 10.1145/1989323.1989438

Brownlee, J. ,2016. *Supervised and unsupervised learning algorithms.* Available at: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms. [Accessed on May 6 2019]

Brownlee, J., 2020. *4 types of classification tasks in machine leaning*. Available at: https://machinelearningmastery.com/types-of-classification-in-machine-learning. [Accessed on 05/04/2021].

Byrd, W.L & Byrd,T.A. , 2013. Contrasting the Dimensions of Information Quality in their Effects on Healthcare Quality in Hospitals, *46th Hawaii International Conference on System Sciences*, Wailea, Maui, HI USA, Jan 7 – 10, 2013.

Caballero, I., Serrano, M. & Piattinni, M., 2014. A data quality in Use model for Big Data. *Indulska M., Purao S. (eds) Advances in Conceptual Modeling. ER 2014. Lecture Notes in Computer Science*, vol 8823. Springer, Cham. https://doi.org/10.1007/978-3-319-12256-4_7.

Cai, L. & Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14(2). DOI: http://doi.org/10.5334/dsj-2015-002

Cappiello,C., Samá,W. & Vitali, M., 2018. Quality awareness for a Successful Big Data Exploitation. *Proceedings of the 22nd International Database Engineering & Applications Symposium* (*IDEAS 2018*). Association for Computing Machinery, New York, NY, USA, 37–44.
DOI:https://doi.org/10.1145/3216122.3216124

Casakin, H. & Singh, V., 2019. Insights from a Latent Semantic Analysis of Patterns in Design Expertise: Implications for Education. *Educ. Sci*. 2019, 9(3), 208; https://doi.org/10.3390/educsci9030208

Chen X.,Cai Y., Liu Q & Chen L., 2018. Nonconvex lp-Norm Regularized Sparse Self-Representation for Traffic Sensor Data Recovery, *IEEE Access*. *10.1109/ACCESS.2018.2832043*

Christen, P., 2008. Febrl – An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* ,August 2008 Pages 1065–1068.https://doi.org/10.1145/1401890.1402020.

Chu, X., Ilyas, I.F. & Papotti, P., 2013. Holistic data cleaning: Putting violations into context, *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, Brisbane, QLD, Australia, pp. 458-469, doi: 10.1109/ICDE.2013.6544847.

Chu, X., Ilyas, I. F., Papotti, P. & Ye, Y., 2014. RULEMINER: Data Quality Rules Discovery. *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA, pp. 1222-1225, doi: 10.1109/ICDE.2014.6816746.

Chu, X., Morcos,J.,Ilyab,I.F., Ouzzani,M., Papotti, P., Tang,N. & Ye,Y. 2015. KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing. *Proceedings of the VLDB Endowment*, Vol. 8, No. 12, 1952–1955. DOI:https://doi.org/10.14778/2824032.2824109

Cichy, C. & Rass, S., 2019. An Overview of Data Quality Frameworks, *IEEE Access*, vol. 7, pp. 24634-24648, doi: 10.1109/ACCESS.2019.2899751.

CIHI, 2015. *Canadian Institute for Health Information.* [Online]
Available at: https://www.cihi.ca/en/data-and-standards/data-quality
[Accessed 01 October 2015].

Corrales,D.C., Ledezma,A. & Corrales, J.C, 2018. From Theory to Practice: A Data Quality Framework for Classification Tasks. *Symmetry*, 10, 248; doi:10.3390/sym10070248

Craswell, A., Moxham, L., Broadbent, M., 2016. Does use of computer technology for perinatal data collection influence data quality? *Health Inform. J.* , *22*, 293–303, doi:10.1177/1460458214556372

Cure, O., 2012. Improving the data quality of drug databases using conditional dependencies and ontologies. *ACM J. Data Inform. Qual.*, *4*, 3, https://doi.org/10.1145/2378016.2378019

Dave, M. & Gianey, K.K., 2016. Analysis of big data for data-intensive applications, *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Jaipur, pp. 1-6, doi: 10.1109/ICRAIE.2016.7939551.

Deerwester,S.,Dumais,S.,Furnas,G.,Landauer,T. & Harshman, R., 1990. Indexing by Latent Semantic Analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*. 41(6):391-407.

Demchenko, Y., de Laat, C. & Membrey, P., 2014. Defining architecture components of the Big Data Ecosystem, *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 104-112, doi: 10.1109/CTS.2014.6867550.

Dong, W., Fong, D.Y.T., Yoon, Js. et al., 2021. Generative adversarial networks for imputing missing data for big data clinical research. BMC Med Res Methodol 21, 78. https://doi.org/10.1186/s12874-021-01272-3

ENGLISH, L. 1999. *Improving Data Warehouse and Business Information Quality*. Wiley & Sons

Ezzine, I. & Benhlima L., 2018. A study of handling missing data methods for big

data. *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. Marrakech, Morocco, 2018, pp. 498-501, doi: 10.1109/CIST.2018.8596389.

Firmani, D.,Mecella,M.,Scannapieco,M. & Batini, C. ,2016 .On the meaningfulness of "Big Data Quality", *Data Science Engineering Journal*, 1(1):6–20, DOI 10.1007/s41019-015-0004-7

García, S., Ramírez-Gallego, S., Luengo, J. *et al.*, 2016. Big data preprocessing: methods and prospects. *Big Data Anal* **1,** 9. https://doi.org/10.1186/s41044-016-0014-0

Gareth, J. ,2013. An Introduction to Statistical Learning: with Applications in R. *Springer*. p. 176. *ISBN 978-1461471370*.

Geisler, S., Weber, S. & Quix, C., 2011. ONTOLOGY-BASED DATA QUALITY FRAMEWORK FOR DATA STREAMS. *Journal of Data and Information Quality*, October 2016 Article No.: 18 https://doi.org/10.1145/2968332.

George, L.E. & Birla, L., 2018. A Study of Topic Modeling Methods. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 109-113, doi: 10.1109/ICCONS.2018.8663152.

Giarizzo-Wilson, S.,2011. Pre-charting patient care information. *AORN Journal*, Volume 94, Issue 6, https://doi.org/10.1016/j.aorn.2011.09.011

Gibson,N. 1997. *Measuring the quality of patient data with particular reference to data accuracy*, University of Keele, UK.

Ginsberg, J., Mohebbi, M., Patel, R. *et al.*, 2009. Detecting influenza epidemics using search engine query data. *Nature* **457,** 1012–1014. https://doi.org/10.1038/nature07634

Glowalla, P., Balazy, P., Basten, D. & Sunyaev, A., 2014. Process-Driven Data Quality Management -- An Application of the Combined Conceptual Life Cycle Model, *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, pp. 4700-4709, doi: 10.1109/HICSS.2014.575.

Gluck,J., 2020. I*t's All About the Data in 2020 and Beyond,* Available at: https://healthtechmagazine.net/article/2020/01/its-all-about-data-2020-and-beyond. [Accessed on July 04 2020]

Gnesotto, R., DeVogli, R., 2003. *Health Monitoring Systems in Europe: Structures and Processes. EC Health Monitoring Programme.* Available at: http://europa.eu.int/comm/health/ph_projects/2001/monitoring/fp_monitoring_2001_frep_13_en.pdf. [Accessed on 10 June 2016]

Handler, D. J., 2012. *Small Data-Thinking Kills Big Data-Aspirations.* [Online] Available at: http://www.wired.com/insights/2013/01/small-data-thinking-kills-big-data-aspirations. [Accessed 29 December 2015].

Hariharakrishnan, J., Mohanavalli,, S. & Sundhara Kumar, K.B., 2017. Survey of pre-processing techniques for mining big data, *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, Chennai, India, pp. 1-5, doi: 10.1109/ICCCSP.2017.7944072.

Harada, N., Yamashita, K., Motomura, Y. & Kano, Y., 2020. Applying Statistical Approach to Topic Analysis for more Comprehensive and Appropriate Modeling. *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, pp. 13-18, doi: 10.1109/CDMA47397.2020.00008.

HDK, 2014. *List of Different Types of Health Data.* [Online]
Available at: http://www.healthdataknowledge.com/list-of-different-types-of-health-data/
[Accessed 23 August 2016].

Helbing, D., Balietti, S., 2011. From social data mining to forecasting socio-economic crises. *Eur. Phys. J. Spec. Top.* **195,** 3. https://doi.org/10.1140/epjst/e2011-01401-8

Hermans, F., 2009. *Data Fusion Based on distributed quality estimations in wireless sensor networks,* Freie Universitat Berlin.

Hima, P. K. et al., 2011. Data Cleansing techniques for Large Enterprise datasets. *2011 Annual SRII Global Conference*, San Jose, CA, USA, 2011, pp. 135-144, doi: 10.1109/SRII.2011.26

Homayouni, H. Ghosh, S. & Ray, I., 2019. ADQuaTe: An Automated Data Quality Test Approach for Constraint Discovery and Fault Detection. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, Los Angeles, CA, USA, 2019, pp. 61-68, doi: 10.1109/IRI.2019.00023.

HRSA (2019). *Data Explorer.* Available at: https://data.hrsa.gov/tools/data-explorer. [Accessed on July 17 2019]

Huang, H., Stvilia, B. & Bass, H., 2012. Prioritization of Data Quality Dimensions and Skills Requirements in Genome Annotation Work. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY,* Vol 63, Issue 1, https://doi.org/10.1002/asi.21652

Huser,V., Kahn,M.,Brown,J. & Gouripeddi,R., 2018. Methods for examining data quality in healthcare integrated data repositories**.** *Pacific Symposium on Biocomputing 2018,* pp 628-633, doi: https://doi.org/10.1142/9789813235533_0059

Jacke, C.O.; kalder, M.; Wagner, U.; Albert, U., 2012. Valid comparisons and decisions based on clinical registers and population based cohort studies: Assessing the accuracy, completeness and epidemiological relevance of a breast cancer query database. *BMC Res. Notes, 5,* 700

Jesmeen, M.Z.H, Hossen,J., Sayeed,S.,Ho,C.K.,Tawsif,K.,Rahman,A. & Arif, E.M.H, 2018. A Survey on Cleaning Dirty Data Using Machine Learning Paradigm for Big Data Analytics. *Indonesian Journal of Electrical Engineering and Computer Science***,** Vol. 10, No. 3, June 2018, pp. 1234~1243 ISSN: 332502-4752, DOI: 10.11591/ijeecs.v10.i3.pp1234-1243

Ji,S., Li,Q., Cao, W., Zhang,P. & Muccini, H., 2020. *Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review.* Appl. Sci.2020,10, 8052; doi:10.3390/app10228052

Jones,k.,Zenk,S.,Tarlov,E.,Powell,L.,Matthews,S. & Horoi,I., 2017. A step-by-step approach to improve data quality when using commercial business lists to characterize retail food environments. *BMC Res Notes,* 10:35 DOI 10.1186/s13104-016-2355-1

Juneja, A. & Das, N., 2019. Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application, *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 559-563, doi: 10.1109/COMITCon.2019.8862267.

Karloff, H., Suri,S. & Vassilvitskii,S.,2010. A model of computation of MapReduce, *Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms* , pp 938-948, doi: 10.1137/1.9781611973075.76

Katerattanakul, P. & Siau, K., 1999. Measuring information quality of web sites: Development of an instrument. *ICIS 1999 Proceedings*. 25. https://aisel.aisnet.org/icis1999/25.

Khan, M.; Raebel, M.A.; Glanz, J.M.; Riedlinger, K.; Steiner, J., 2012. A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research. *Med. Care*, doi:10.1097/MLR.0b013e318257dd67

Khayyat, Z. et al., 2015. BigDansing: A system for Big Data Cleansing. *SIGMOD '15: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. May 2015 Pages 1215–1230.https://doi.org/10.1145/2723372.2747646.

Kim, W., 2002. On three major holes in Data Warehousing Today. *Journal of Object Technology.* Vol. 1, no. 4, September-October 2002

Kim,J.,Tae, D. & Seok, J.,2020.A Survey of Missing Data Imputation Using Generative Adversarial Networks. *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 454-456, doi: 10.1109/ICAIIC48513.2020.9065044.

King, R., Feng, C. & Sutherland, A., 1995. Statlog: comparison of classification algorithms on large real-world problems, *Applied Artificial Intelligence*, 9:3, 289-333, DOI: 10.1080/08839519508945477.

Kitchin, R., Lauriault, T.P., 2015. Small data in the era of big data. *GeoJournal* **80,** 463–475. https://doi.org/10.1007/s10708-014-9601-7

Kitchin, R. & McArdle, G., 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. doi:10.1177/2053951716631130

Krishnan, S., Wang, J., Wu, E., Franklin, M.J. & Goldberg, K., 2015. Activeclean: Interactive data cleaning while learning convex loss models. Accessible at: Arxiv:http:// arxiv.org/ pdf/1601.03797.pdf

Krishnan, S., Wu, E., Franklin, M.J & Goldberg,K., 2017. BoostClean: Automated Error Detection and Repair for Machine Learning. Accessible at: https://www.groundai.com/project/boostclean-automated-error-detection-and-repair-for-machine-learning/1. [Accessed on 15/04/2020]

Kulkarni, S.S., Apte, U.M. & Evangelopoulos,N.E. , 2014. The use of Latent Semantic Analysis in Operations Management Research. *Journal of Decision Sciences Institute*, Volume 45, No. 5.

Laranjeiro,N., Soydemir,S. & Bernardino,J. (2015). A survey on data quality: Classifying Poor Data. *The 21st IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2015),* Zhangjiajie, China, 2015, pp. 179-188, doi: 10.1109/PRDC.2015.41

Labouseur,I. & Matheus, C., 2017. An Introduction to Dynamic Data Quality Challenges. *J. Data and Information Quality* 8, 2, Article 6 (February 2017), 3 pages. DOI:https://doi.org/10.1145/2998575

Lakshen, G. & Vranes, S. , 2016. Big Data and Quality: A Literature Review. *24th Telecommunications forum TELFOR 2016*. Serbia, Belgrade, November 22-23, 2016

Landauer, T. K., Foltz, P. W., & Laham, D.,1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**, 259-284, doi: https://doi.org/10.1080/01638539809545028

Langley, J., Stephenson, S., Thorpe, C. & Davie, G., 2006. Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges. *Injury Prevention,* Volume 12, pp. 58-61.

LEE, Y.W., STRONG, D. M., KAHN, B. K. & WANG, R. Y., 2002. AIMQ: A methodology for information quality assessment. *Inform. Manage. 40*, 2, 133–460

Leon, A.; Reyes, J.; Burriel, V.; Valverde, F., 2016. Data Quality Problems when Integrating Genomic Information. *Proceedings of the ER 2016 Workshops, LNCS 9975*, Gifu, Japan, 14–17 November 2016; pp. 173–182, doi:10.1007/978-3-319-47717-6 15

Lee, H. S. & Haider A., 2013. Identifying Relationships of Information Quality Dimensions. *2013 Proceedings of PICMET '13: Technology Management for Emerging Technologies*, San Jose, California, 28th J^{ul}y – 01st August 2013

Lima, C.R.; Schramm, J.M.; Coeli, C.M.; da Silva, M.E. , 2009. Review of data quality dimensions and applied methods in the evaluation of health information systems. *Cad. Saúde Pública*, *25*, 2095–2109

Li, X., Shi, Y., Li, J. & Zhang, P., 2007. Data mining consulting improve data quality. *Data Science Journal*, Issue 6, https://doi.org/10.2481/dsj.6.S658

Lin, T., Yang, C. & Chiang, I., 2014. Improvement of prognostic models for ESRD mortality by the bootstrap method with random hot deck imputation, *2014 IEEE International Conference on Granular Computing (GrC)*, Noboribetsu, pp. 166-169.doi: 10.1109/GRC.2014.6982828

Lin, X.,Lin, P. & Huang, P., 2015. Modeling the Task of Google MapReduce Workload, *15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, **DOI:** 10.1109/CCGrid.2015.104

Loh, M.J & Dasu T., 2012. Effect of Data Repair on Mining Network Streams, *2012 IEEE 12th International Conference on Data Mining Workshops*, Brussels, Belgium, 2012, pp. 226-233, doi: 10.1109/ICDMW.2012.125.

LOSHIN, D. 2004. Enterprise Knowledge Management - The Data Quality Approach. *Series in Data Management Systems,* Morgan Kaufmann, chapter 4.

Loshin, D., 2014. *Understanding Big Data Quality for Maximum Information Usability*, White Paper, SASA Institute Inc, **2014**

Mahdavi, M., Neutatz, F., Visengeriyeva, L. & Abedjan, Z., 2019. Towards Automated Data Cleaning Workflows, *Conf. on 'Lernen, Wissen, Daten, Analysen*, Berlin, Germany pp. 10–19

Mahdavi, M. & Abedjan, Z., 2021. Semi-Supervised Data Cleaning with Raha and Baran, *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*, January 10-13, 2021, Chaminade, USA

Majed, F., 2016. *Machine Learning Ensemble Method for discovering knowledge from Big Data*, PhD, University of East Anglia.

Malik, P., 2013. Governing Big Data: Principles and Practices. *IBM Journal of Research and Development*, Vol. 57, no. 3/4, pp. 1:1-1:13, May-July 2013, doi: 10.1147/JRD.2013.2241359

Marr, B., 2015. *How Big Data is Changing Healthcare.* [Online]
Available at: www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare.
[Accessed 01 October 2015].

Maktoubian, J., 2019. MongoDB, *ISIM*, DOI: 10.13140/RG.2.2.16032.43520

Maxwell-Downing, D., 2006. Data quality and the electronic health record (EHR). *AORN J.*, Volume 94, Issue 6, https://doi.org/10.1016/j.aorn.2011.09.011

Micic, N., Neagu, D., Campean, F. & Zadeh, H. , 2017. Towards a Data Quality Framework for Heterogeneous Data. *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom)and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Exeter, UK, 2017, pp. 155-162, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.28.

Miller, J.E., Gaboda, D., Davis, D., 2001. Early childhood chronic illness: Comparability of maternal reports and medical records. *National Center for Health Statistics*. Vital Health Stat 2(131).

Mishra, A., 2018. *Metrics to evaluate your machine learning algorithm.* Available at: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. [Accessed on May 08 2019]

Mohmad, N. Hassan, N., Samy, G., Aziz,N. Maaro.p,N. & Bakar N., 2020. *A Review of Factors Influencing Patient Readmission Based on Data Quality Dimension Model.* 2020 8th International Conference on Information Technology and Multimedia (ICIMU), 24 - 26 Aug 2020, Selangor, Malaysia. IEEE

Müller, H. & Freytag, J., 2003. *Problems, Methods, and Challenges in Comprehensive Data Cleansing*, Technical Report HUB-IB-164, Humboldt University Berlin, 2003

Nag, A., 2019. *Unsupervised outlier detection in text corpus using Deep Learning.* Available at: https://medium.datadriveninvestor.com/unsupervised-outlier-detection-in-text-corpus-using-deep-learning-41d4284a04c8. [Accessed on December 03 2021]

Negri, R. G., Sant'Anna, S. J. S. & Dutra, L. V., 2011. Semi Supervised Remote Sensing Image Classification Methods. *011 IEEE International Geoscience and Remote Sensing Symposium*, Vancouver, BC, Canada, 2011, pp. 2939-2942, doi: 10.1109/IGARSS.2011.6049831.

Nystrom, M., Andersson,R.,Holmqvist,K. & Weijer,J. , 2013. The influence of calibration method and eye physiology on eyetracking data quality, *Behavioral Research*, vol 45, pp 272 – 288. DOI 10.3758/s13428-012-0247-4

Obhyung Kwon, N. L. S., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management,* pp. 387-394.

Oliveira, P., Rodrigues, F., Henriques, P. & Galhardas, H., 2005. A Taxonomy of Data Quality Problems. *2nd Int. Workshop on Data and Information Quality*, pp 219-233, Porto.

Onyeabor, G.A. & Ta'a A., 2019. A Model for Addressing Quality Issues in Big Data. *Saeed F., Gazem N., Mohammed F., Busalim A. (eds) Recent Trends in Data Science and Soft Computing. IRICT 2018. Advances in Intelligent Systems and Computing*, vol 843. Springer, Cham. https://doi.org/10.1007/978-3-319-99007-1_7

O'Reilly, G., Gabbe, B., Moore, L. & Cameron, P., 2016. Classifying, measuring and improving the quality of data in trauma registries: A review of literature. *Injury, Int. J. Care Injured,* Volume 47, pp. 559-567, https://doi.org/10.1016/j.injury.2016.01.007.

Pam, W. 2014. A pilot ontology for a large, diverse set of national health service healthcare quality indicators. (Unpublished Doctoral thesis, City University London)

Panahy, P. et al., 2013. Discovering dependencies among data quality dimensions: A validation of instrument. *Journal of applied sciences,* 13(1), pp. 95-102.DOI: 10.3923/ jas.2013.95.102

Pandove, D., Goel, S. and Rani, R., 2018. *Systematic Review of Clustering High-Dimensional andLarge Datasets*.ACM Trans. Knowl. Discov. Data.12, 2, Article 16 (January 2018), 68 pages. https://doi.org/10.1145/3132088

Pang, L., Lan, Y., Guo, J.,Xu, J.,Xu,J. & Cheng, X., 2017. DeepRank: A New Deep Architecture for Relevance Ranking in Information Retrieval. *CIKM'17*, November 6–10, 2017, Singapore. DOI: hŠ p://dx.doi.org/10.1145/3132847.3132914

Pasteels, J.M, 2013. *Review of best practice methodologies for imputing and harmonising data in cross-country datasets,* ILO internal report

Pinto M. Data representation factors and dimensions from the quality function deployment (QFD) perspective. *Journal of Information Science*. 2006;32(2):116-130. doi:10.1177/0165551506062325

Pipino, L., Yang, L. & Wang, R., 2002. Data Quality Assessment. *Communications of the ACM.*Vol. 45, Issue 4. https://doi.org/10.1145/505248.506010

Pittsburgh health data alliance, 2016. *Three Pittsburgh Institutions. One Goal* [Online] Available at: https://healthdataalliance.com/ [Accessed 12 March 2018]

Qahtan, A., Ouzzani,M., Elmagarmid, A. and Tang,N. 2018. FAHES: A Robust Disguised Missing Values Detector. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (*KDD '18*). Association for Computing Machinery, New York, NY, USA, 2100–2109. DOI:https://doi.org/10.1145/3219819.3220109

Raghupathi, V. & Raghupathi, W., 2014. Big data analytics in healthcare: promise and. *Health information science and systems,* 2(3), https://doi.org/10.1186/2047-2501-2-3.

Rahimi,A.,Taggart,J.,Parameswaran,N.,Yu,H., Ray,P.K. & Liaw,S., 2016. *Development of a Methodological Approach for Data Quality Ontology in Diabetes Management.* IGI global. DOI: 10.4018/978-1-4666-8756-1.ch023

Rahm, E. & Do, H., 2000. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*, pp. 3-13.

Rahman, G., Islam, Z.,Bossomaier, T. & Gao, J., 2012. CAIRAD: A Co-appearance based Analysis for Incorrect Records and Attribute-values Detection, *WCCI 2012 IEEE World Congress on Computational Intelligence* June, 10-15, 2012 - Brisbane, Australia

Raman, V. & Hellerstein, J., 2001. Potter's Wheel: An Interactive Data Cleaning System. *Proceedings of the 27th VLDB Conference*, Roma, Italy, 2001.

Rao, D., Gudivada,V. & Raghavan, V., 2015. Data Quality issues in Big Data, *2015 IEEE International Conference on Big Data*, Oct 29 - Nov 01, Santa Clara, CA, USA, pp. 2654-2660, doi: 10.1109/BigData.2015.7364065.

Rekatsinas, T., Xu, C., Ilyas, F. & Re,C., 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference**,** *Proceedings of the VLDB Endowment*, Vol. 10, No. 11

Rhodegero, J. The use of big data in manual physiotherapy. *Man. Ther*. **2014**, *19*, 509–510, doi: 10.1016/j.math.2014.10.014

Ridzuan, F., & Zainon, W., 2019. A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science* 161 (2019) 731–738, https://doi.org/10.1016/j.procs.2019.11.177

Salati M, Falcoz P-E, Decaluwe H, Rocco G, Van Raemdonck D, Varela G et al., 2016. The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases. *Eur J Cardiothorac Surg*;49:1470–5

Sackett, D., Rosenberg,W., Gray, J.A., Haynes,R. & Richardson, W., 1996. Evidence based medicine: what it is and what it isn't. *BMJ,* 312(71). doi: https://doi.org/10.1136/bmj.312.7023.71

Sadineni, P.K, 2020. *Developing a Model to Enhance the Quality of Health Informatics using Big Data*. Proceedings of the Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)IEEE Xplore Part Number:CFP20OSV-ART; ISBN: 978-1-7281-5464-0

Saha, B. & Srivastava, D., 2014. Data Quality: The other face of Big Data. *014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA, 2014, pp. 1294-1297, doi: 10.1109/ICDE.2014.6816764.

Schafer, J. & Graham J., 2002. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Seppala, T., 2015. *Jawbone's Up3 finally ships on April 20th, but don't take it swimming.* [Online] Available at: https://www.engadget.com/2015/04/10/jawbone-up3-shipping-april-20th [Accessed 12 March 2018]

Serhani, M.A.,Kassabi,H.T,Taleb,I. & Nujum,A., 2016. An Hybrid Approach to Quality Evaluation Across Big Data Value Chain, *2016 IEEE International Congress on Big Data*, 2016, June 27 - July 2, 2016, San Francisco, USA. doi: 10.1109/BigDataCongress.2016.65.

Srivastava, D., Scannapieco, M. and Redman, T. (2019). *Ensuring High-Quality Private Data for Responsible Data Science: Vision and Challenges*.J. Data and Information Quality11, 1, Article 1 (January2019), 9 pages. https://doi.org/10.1145/3287168

Shi, W. et al., 2015. Improving Power Grid Monitoring Data Quality: An Efficient Machine Learning Framework for Missing Data Prediction. *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, New York, NY, USA, pp. 417-422, doi: 10.1109/HPCC-CSS-ICESS.2015.16.

Sidi, F. Jabar,M.,Ibrahim,H. & Mustapha, A., 2012. *Data Quality: A survey of data quality dimensions. 2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia, pp. 300-304, doi: 10.1109/InfRKM.2012.6204995.

Smyth, D., 2004. *The Health Information Strategy*. Presentation to the Irish Forum for Health Informatics, Dublin, Oct 7, 2004. Available at: http://www.ifhi.ie. [Accessed on 10 June 2016]

Soares, S., 2012. Big Data quality. In: *Big Data Governance: An emerging imperative*. MC Press, pp. 101-112.

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S. & Yahia, S., 2019. Data quality in ETL process: A preliminary study. *Procedia Computer Science* 159 (2019) 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Spengler H., Gatz, I., Kohlmayer, F., Kuhn, K. & Prasser F., 2020. Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA, 2020, pp. 415-420, doi: 10.1109/CBMS49503.2020.00085.

Sporleder, C., Erp, M.V., Porcelijn, T., & Bosch, A.V. ,2006. Spotting The 'Odd-One-Out': Data-Driven Error Detection And Correction In Textual Databases. Accessible at: https://www.aclweb.org/anthology/W06-2206.pdf. [Accessed on 24 April 2018]

Sukumar R., Ramachandran N., and Ferrell R. K., 2015. 'Big Data' in health care: How good is it?' *International Journal of Health Care Quality Assurance*, 2-9

Sushovan, D., Yuheng, H., Yi, C. & Subbarao, K., 2014. BayesWipe: A Multimodal System for Data Cleaning and Consistent Query Answering on Structured BigData. *2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2014, pp. 15-24, doi: 10.1109/BigData.2014.7004207.

Syntelli Marketing, 2020. *The Role of Big Data in Preventing Healthcare Fraud and Waste.* [Online] Available at: https://www.syntelli.com/role-big-data-preventing-healthcare-fraud-waste [Accessed 03 February 2021].

Taleb, I., Dssouli, R. & Serhani, M.A, 2015. Big Data Pre-processing: A Quality Framework, *2015 IEEE International Congress on Big Data*, New York, NY, USA, pp. 191-198, doi: 10.1109/BigDataCongress.2015.35.

Taleb, I., Kassabi, H., Serhani, M.A., Dssouli, R. & Bouhaddiou C., 2016. Big Data Quality: A Quality Dimensions Evaluation, *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122

Taleb,I., Serhani, M.A. & Dssouli, R., 2018. Big Data Quality: A Survey, *2018 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, USA, pp. 166-173, doi: 10.1109/BigDataCongress.2018.00029.

Tett, G., 2013. *Privacy fears are holding back a healthcare revolution*, London, UK Financial Times

Thusoo, A., Sarma, J.S. & Jain N.., 2010. Hive - A Petabyte Scale Data Warehouse Using Hadoop. *Proceedings of the 26th International Conference on Data Engineering*, ICDE 2010, March 1-6, 2010, Long Beach, California, USA. DOI: 10.1109/ICDE.2010.5447738

Todoran, I.-G., Lecornu, L., Kenchaf, A. & Le Caillac, J. M., 2015. A Methodology to Evaluate Important Dimensions of Information quality in systems. *Journal of Data and Information Quality*, Vol 6, Issue 2-3, https://doi.org/10.1145/2744205.

Torraco R., 2005. Writing Integrative Literature Reviews: Guidelines and Examples. *Hum. Resour. Dev. Rev.*, *4*, 356–367, doi:10.1177/1534484305278283

Tran, C., 2018. *Evolutionary Machine Learning for Classification with incomplete data*, PhD, Victoria University of Wellington, New Zealand

UNECE, 2014. *A suggested framework for the quality of Big Data*.Available at: http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Fr

amework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2. [Accessed 24 Sep 2016]

Varshney, K. R., Wei, D., Ramamurthy, K. N. & Mojsilovic, A., 2015. Data Challenges in Disease Response: The 2014 Ebola Outbreak and beyond. *ACM J. Data Inf. Qual,* Volume 6, pp. 2-3. https://doi.org/10.1145/2742550.

Vattulainen, M., 2015. Improving the Predictive Power of Business Performance measurement systems by Constructed Data Quality Features?Five cases. Perner P. (eds) Advances in Data Mining: Applications and Theoretical Aspects. ICDM 2015. Lecture Notes in Computer Science, vol 9165. Springer, Cham. https://doi.org/10.1007/978-3-319-20910-4_1.

Vaughan J.., 2015. *data quality*. [Online]
Available at: http://searchdatamanagement.techtarget.com/definition/data-quality
[Accessed 4 May 2015]

Vetro, A., Canova, L., Torchiano, M. & Iemma, R., 2016. Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly,* Vol. 33, Issue 2, pp. 325-337. https://doi.org/10.1016/j.giq.2016.02.001.

Vohra D. ,2016. Apache Hive. *Practical Hadoop Ecosystem*. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2199-0_3

Wahner, K., 2014. *Real-Time Stream Processing as Game Changer in a Big Data World with Hadoop and Data Warehouse.* [Online] Available at: https://www.infoq.com/articles/stream-processing-hadoop. [Accessed 17 July 2017]

Wang, R. & Strong, D., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* 12(4), pp. 5 - 33. DOI: 10.1080/07421222.1996.11518099.

WANG, R. 1998. A product perspective on total data quality management. *Comm. ACM* 41, 2, 58–65. DOI:https://doi.org/10.1145/269012.269022

Wang, H. Li, M. Bu, Y., Li,J., Gao,H. and Zhang,J. ,2014. Cleanix: A Big Data Cleaning Parfait. *CIKM'14,* November 3–7, 2014, Shanghai, China. ACM 978-1-4503-2598-1/14/11.

Weber, J.; Price, M.; Davies, I., 2015. Data Quality by Contract—Towards an Architectural View for Data Quality in Health Information Systems. *Proceedings of the Conference on Artificial Intelligence in Medicine in Europe*, Pavia, Italy, 17–20 June 2015; pp. 143–157, doi:10.1007/978-3-319-26585-8_10

Weiskopf, N. G. & Chunhua, W., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of Americam Medical Information Association,* Issue 20, pp. 144-151. https://doi.org/10.1136/amiajnl-2011-000681.

Whittemore, R.; Knafl, K. , 2005. The integrative review: Updated methodology. *J. Adv. Nurs. ,5,* 546–553

Wu, X. & Zhu, X., 2008. Mining with noise knowledge: Error-Aware Data Mining, IEEE *TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 38, NO. 4,* vol. 38, no. 4, pp. 917-932, July 2008, doi: 10.1109/TSMCA.2008.923034.

Xiao Y, Bochner AF,Makunike B, et al., 2017. Challenges in data quality: the influence of data quality

assessments on data availability and completeness in a voluntary medical male circumcision programme in Zimbabwe. *BMJ Open*. doi:10.1136/bmjopen-2016013562

Xiaolan, W., Xin Luna, D. & Alexandra, M., 2015. Data X-Ray: A diagnostic tool for data errors. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '15*). Association for Computing Machinery, New York, NY, USA, 1231–1245. DOI:https://doi.org/10.1145/2723372.2750549.

Xu, B. et al., 2014. Ubiquitous Data Accessing Method in IoT-Based Information System for Emergency Medical Services. *EEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1578-1586, May 2014, doi: 10.1109/TII.2014.2306382

Yakout, M., Elmagarmid, A. K. & Neville, J., 2010. Ranking for data repairs. *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, Long Beach, CA, USA, 2010, pp. 23-28, doi: 10.1109/ICDEW.2010.5452767.

Yakout, M., Berti-Équille, L. & Elmagarmid. A., 2013. Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes. *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (*SIGMOD '13*). Association for Computing Machinery, New York, NY, USA, 553–564. DOI:https://doi.org/10.1145/2463676.2463706

Yeh, P. Z. & Puri, C. A., 2010. An efficient and robust approach for discovering Data Quality Rules. *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, Arras, France, pp. 248-255, doi: 10.1109/ICTAI.2010.43

Yesha, Y., Janeja, V., Rishe, N. & Yesha, Y., 2014. Personalized Decision Support System to Enhance Evidence Based Medicine through Big Data Analytics. *2014 IEEE International Conference on Healthcare Informatics*, Verona, Italy, pp. 376-376, doi: 10.1109/ICHI.2014.71.

Yinghao, H., Yi Lu,M. & Yao, G., 2013. Automotive diagnosis typo correction using domain knowledge and machine learning. *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Singapore, pp. 267-274, doi: 10.1109/CIDM.2013.6597246.

Yoon, J., Jordon,J. & Schaar, M., 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. *Proceedings of the35thInternational Conference on MachineLearning, Stockholm, Sweden, PMLR 80*

Youngwa Lee, K. K. R. L., 2003. The Technology acceptance model:Past, present and future. *Communication for the association of information systems.* Vol. 12 , Article 50. DOI: 10.17705/1CAIS.01250

Yu, W., Zhu, W., Liu,G.,Kan,B., Zhao, T. & Liu, H. , 2017. Cluster-based Best Match Scanning for Large-Scale Missing Data Imputation. *2017 3rd International Conference on Big Data Computing and Communications*. DOI 10.1109/BIGCOM.2017.48

Yuan, Y., Li, S., Zhang, X. & Sun, J., 2018. A comparative analysis of SVN, naïve bayes and GBDT for data faults detection in WSNs. *2018 IEEE International Conference on Software Quality, Reliability and Security Companion*, Lisbon, Portugal, 2018, pp. 394-399, doi: 10.1109/QRS-C.2018.00075.

Zillner, S., Oberkampf, H., Bretschneider, C. & Amrapali, Z., 2014. Towards a Technology Roadmap for Big Data Applications in the healthcare domain. *Proceedings of the 2014 IEEE 15th International*

Conference on Information Reuse and Integration (IEEE IRI 2014), Redwood City, CA, USA, 2014, pp. 291-296, doi: 10.1109/IRI.2014.7051902.

Zolfaghar, K. et al., 2013. Big data solutions for predicting risk-of-readmission for congestive heart failure patients. *2013 IEEE International Conference on Big Data*, Silicon Valley, CA, USA, 2013, pp. 64-71, doi: 10.1109/BigData.2013.6691760.

Zhang, C., Qin, Y., Zhu, X., Zhang, J. & Zhang, S.,2006. Clustering-based Missing Value Imputation for Data Preprocessing, *2006 4th IEEE International Conference on Industrial Informatics*, Singapore, 2006, pp. 1081-1086.doi: 10.1109/INDIN.2006.275767

Zhang, P., Xiong, F., Gao, J. & Wang, J., 2017. Data quality in big data processing: Issues, solutions and open problems. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, San Francisco, CA, USA, 2017, pp. 1-7, doi: 10.1109/UIC-ATC.2017.8397554

Appendices

Appendix 1: Imputation scripts developed in python 2.7 and 3.8

Isotonic Regression

```
import pandas as pd
#from sklearn.linear_model import LinearRegression
from sklearn.isotonic import IsotonicRegression
from google.cloud import storage
from io import BytesIO
from time import perf_counter
import numpy as np
from sklearn import metrics
client = storage.Client()
bucket_name = "health_ds"
file_name = "BRFSS.csv"
bucket = client.get_bucket(bucket_name)
blob = bucket.get_blob(file_name)
content = blob.download_as_string()
df = pd.read_csv(BytesIO(content),usecols = ['Sample_Size','Confidence_limit_Low'])
linreg = IsotonicRegression()
data = df[['Sample_Size','Confidence_limit_Low']]
original_DS = df.Confidence_limit_Low
x_train = data[data['Confidence_limit_Low'].notnull()].drop('Confidence_limit_Low', axis= 1)
y_train = data[data['Confidence_limit_Low'].notnull()]['Confidence_limit_Low']
x_test = data[data['Confidence_limit_Low'].isnull()].drop('Confidence_limit_Low', axis=1)
y_test = data[data['Confidence_limit_Low'].isnull()]['Confidence_limit_Low']
#Step-2: Train the machine learning algorithm

linreg.fit(x_train, y_train)

#Step-3: Predict the missing values in the attribute of the test data.

predicted = linreg.predict(x_test)

#Step-4: Let’s obtain the complete dataset by combining with the target attribute.


df.Confidence_limit_Low[df.Confidence_limit_Low.isnull()] = predicted
#print(df.Confidence_limit_Low) #getting the imputed data frame
outliers=[]
def detect_outlier(data_1):

    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)
    count = 0

    for y in data_1:
        z_score= (y - mean_1)/std_1
```

```
        if np.abs(z_score) > threshold:
            #outliers.append(y)
            count = count + 1
    return count

# detecting outliers in original dataset
outlier_datapoints = detect_outlier(original_DS)
#detecting outliers in imputed dataset

time_start = perf_counter()

outlier_datapoints = detect_outlier(df.Confidence_limit_Low)
print(outlier_datapoints)

time_stop = perf_counter()
print("Elapsed time:", time_stop - time_start)
```

Linear regression

```
from sklearn.linear_model import LinearRegression
import numpy as np
from pandas import read_csv

train = read_csv('EHR.csv', usecols = ['Attestation_Year','Payment_Year'])
linreg = LinearRegression()
data = train[['Attestation_Year','Payment_Year']]
original_DS = train.Payment_Year

#Step-1: Split the dataset that contains the missing values and no missing values are test and
train respectively.

x_train = data[data['Payment_Year'].notnull()].drop('Payment_Year', axis= 1)
y_train = data[data['Payment_Year'].notnull()]['Payment_Year']
x_test = data[data['Payment_Year'].isnull()].drop('Payment_Year', axis=1)
y_test = data[data['Payment_Year'].isnull()]['Payment_Year']

#Step-2: Train the machine learning algorithm

linreg.fit(x_train, y_train)

#Step-3: Predict the missing values in the attribute of the test data.

predicted = linreg.predict(x_test)

#Step-4: Let's obtain the complete dataset by combining with the target attribute.

train.Payment_Year[train.Payment_Year.isnull()] = predicted
#print(train.Payment_Year) #getting the imputed data frame

'''
```

```
Section for applying outlier metrics for plausability evaluation
'''

outliers=[]
def detect_outlier(data_1):

    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)
    count = 0

    for y in data_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            #outliers.append(y)
            count = count + 1
    return count

# detecting outliers in original dataset
#outlier_datapoints = detect_outlier(original_DS)
#detecting outliers in imputed dataset
outlier_datapoints = detect_outlier(train.Payment_Year)
print(outlier_datapoints)
```

lp- norm regularisation (SRSp)

```
import numpy as np
from sklearn import linear_model
from pandas import read_csv

train = read_csv('EHR.csv', usecols = ['Attestation_Year','Payment_Year'])
data = train[['Attestation_Year','Payment_Year']]

x_train = data[data['Payment_Year'].notnull()].drop('Payment_Year', axis= 1)
y_train = data[data['Payment_Year'].notnull()]['Payment_Year']
x_test = data[data['Payment_Year'].isnull()].drop('Payment_Year', axis=1)
y_test = data[data['Payment_Year'].isnull()]['Payment_Year']

clf = linear_model.SGDRegressor(penalty='elasticnet',alpha=0.0005,l1_ratio=0.2)
clf.fit(x_train, y_train)
predicted = clf.predict(x_test)
train.Payment_Year[train.Payment_Year.isnull()] = predicted
#print(train.Payment_Year)

'''
Section for applying outlier metrics for plausability evaluation
By using z-score
'''

outliers=[]
```

```
def detect_outlier(data_1):

    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)
    count = 0

    for y in data_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            #outliers.append(y)
            count = count + 1
    return count


#detecting outliers in imputed dataset
outlier_datapoints = detect_outlier(train.Payment_Year)
print(outlier_datapoints)
```

Cluster-Based Best Match Scanning (CBMS)

```
from sklearn.cluster import KMeans
import numpy as np
from sklearn.model_selection import train_test_split
from pandas import read_csv
from sklearn.neighbors import KNeighborsClassifier

train = read_csv('EHR.csv', usecols = ['Attestation_Year','Payment_Year'])
data = train[['Attestation_Year','Payment_Year']]

x_train = data[data['Payment_Year'].notnull()].drop('Payment_Year', axis= 1)
y_train = data[data['Payment_Year'].notnull()]['Payment_Year']
x_test = data[data['Payment_Year'].isnull()].drop('Payment_Year', axis=1)
y_test = data[data['Payment_Year'].isnull()]['Payment_Year']


kmeans = KMeans(n_clusters=6, random_state=0).fit(x_train,y_train)


#need to get 10 list of clusters

labels = kmeans.labels_

for i in range(6):
    #for each cluster, apply KNN algo

    A = x_train[(labels == i)]
    B = y_train[(labels == i)]
    c = abs(B.count())
    #print(c)
```

```
   if (c > 2):
      A_train, A_test, B_train, B_test = train_test_split(A, B, test_size=0.3, random_state=1,
stratify=B)

      # Create KNN classifier
      knn = KNeighborsClassifier(n_neighbors = 1, metric = 'correlation')

      # Fit the classifier to the data
      knn.fit(A_train,B_train)
      d = np.array(B_test).reshape(-1,1)
      knn.predict(d)
      print(d)
```

Appendix 2

Extreme value analysis

```
import numpy as np
from pandas import read_csv

data = read_csv('EHR.csv', usecols = ['Product_Certification_Edition_Yr'])

IQR = data.Product_Certification_Edition_Yr.quantile(0.75) -
data.Product_Certification_Edition_Yr.quantile(0.25)
upper_limit = data.Product_Certification_Edition_Yr.quantile(0.75) + (IQR * 1.5)
upper_limit_extreme = data.Product_Certification_Edition_Yr.quantile(0.75) + (IQR * 3)

total = np.float(data.shape[0])
print('Total data: {}'.format(data.Product_Certification_Edition_Yr.shape[0]/total))
print('data over upper limit: {}'.format(data[data.Product_Certification_Edition_Yr >
upper_limit].shape[0]/total))
print('Data over upper limit extreme: {}'.format(data[data.Product_Certification_Edition_Yr >
upper_limit_extreme].shape[0]/total))
```

Text clustering using k means

```
from sklearn.cluster import KMeans
from sklearn.metrics import adjusted_rand_score
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer

from pandas import read_csv

df = read_csv('EHR.csv', usecols = ['Program_Type'])

documents = str(df.dropna().values.tolist()).split(",")

tfidf_vectorizer=TfidfVectorizer(use_idf=True, stop_words = 'english')
```

```
# just send in all your docs here
tfidf_vectorizer_vectors=tfidf_vectorizer.fit_transform(documents)
first_vector_tfidfvectorizer=tfidf_vectorizer_vectors[0]

# place tf-idf values in a pandas data frame
df1 = pd.DataFrame(first_vector_tfidfvectorizer.T.todense(),
index=tfidf_vectorizer.get_feature_names(), columns=["tfidf"])

#with pd.option_context('display.max_rows', None, 'display.max_columns', None):print(df1)
true_k = 1
model = KMeans(n_clusters=true_k, init='k-means++', max_iter=100, n_init=1)
model.fit(tfidf_vectorizer_vectors)

print("Top terms per cluster:")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = tfidf_vectorizer.get_feature_names()
for i in range(true_k):
    print("Cluster %d:" % i),
    for ind in order_centroids[i, :100]:
        print(' %s' % terms[ind]),
    print
```

MLPRegressor

```
from sklearn.cluster import KMeans
from sklearn.metrics import adjusted_rand_score
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from time import perf_counter
from gensim.models.doc2vec import TaggedDocument, Doc2Vec
from gensim.parsing.preprocessing import preprocess_string
from sklearn.base import BaseEstimator
from sklearn import utils as skl_utils
#from tqdm import tqdm

import multiprocessing
import numpy as np

#time_start = time.clock()
time_start = perf_counter()

from pandas import read_csv

#df = read_csv('EHR.csv', usecols = ['Specialty'])
df = read_csv('BRFSS.csv', usecols = ['Class'])

class Doc2VecTransformer(BaseEstimator):

    def __init__(self, vector_size=100, learning_rate=0.02, epochs=20):
        self.learning_rate = learning_rate
```

```
        self.epochs = epochs
        self._model = None
        self.vector_size = vector_size
        self.workers = multiprocessing.cpu_count() - 1

    def fit(self, df_x, df_y=None):
        tagged_x = [TaggedDocument(str(row['Class']).split(), [index]) for index, row in
df_x.iterrows()]
        model = Doc2Vec(documents=tagged_x, vector_size=self.vector_size,
workers=self.workers)

        for epoch in range(self.epochs):
            model.train(skl_utils.shuffle([x for x in tqdm(tagged_x)]),
total_examples=len(tagged_x), epochs=1)
            model.alpha -= self.learning_rate
            model.min_alpha = model.alpha

        self._model = model
        return self

    def transform(self, df_x):
        return np.asmatrix(np.array([self._model.infer_vector(str(row['Class']).split())
                                     for index, row in df_x.iterrows()]))

doc2vec_tr = Doc2VecTransformer(vector_size=300)
doc2vec_tr.fit(df)
doc2vec_vectors = doc2vec_tr.transform(df)

print(doc2vec_vectors)
from sklearn.neural_network import MLPRegressor

auto_encoder = MLPRegressor(hidden_layer_sizes=(600,150, 600, ))
auto_encoder.fit(doc2vec_vectors, doc2vec_vectors)
predicted_vectors = auto_encoder.predict(doc2vec_vectors)
auto_encoder.score(predicted_vectors, doc2vec_vectors)
from scipy.spatial.distance import cosine

def key_consine_similarity(tupple):
    return tupple[1]

def get_computed_similarities(vectors, predicted_vectors, reverse=False):
    data_size = len(title_plot_df)
    cosine_similarities = []
    for i in range(data_size):
        cosine_sim_val = (1 - cosine(vectors[i], predicted_vectors[i]))
        cosine_similarities.append((i, cosine_sim_val))

    return sorted(cosine_similarities, key=key_consine_similarity, reverse=reverse)

def display_top_n(sorted_cosine_similarities, n=5):
```

```
    for i in range(n):
        index, consine_sim_val = sorted_cosine_similarities[i]
        print('Movie Title: ', df.iloc[index, 0])
        print('Cosine Sim Val :', consine_sim_val)
        print('---------------------------------')

print('Top 5 unique classes')
sorted_cosine_similarities = get_computed_similarities(vectors=doc2vec_vectors,
predicted_vectors=predicted_vectors)
display_top_n(sorted_cosine_similarities=sorted_cosine_similarities)
```

GAIN algorithm classes

Main.py

```
# Necessary packages
from __future__ import absolute_import
from __future__ import division
from __future__ import print_function

import argparse
import numpy as np

from data_loader import data_loader
from gain import gain
from utils import rmse_loss


def main (args):

  outliers=[]
  def detect_outlier(data_1):

    threshold=3
    mean_1 = np.mean(data_1)
    std_1 =np.std(data_1)
    count = 0

    for y in data_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            #outliers.append(y)
            count = count + 1
    return count

  data_name = args.data_name
  miss_rate = args.miss_rate

  gain_parameters = {'batch_size': args.batch_size,
                     'hint_rate': args.hint_rate,
```

```
                'alpha': args.alpha,
                'iterations': args.iterations}

  # Load data and introduce missingness
  ori_data_x, miss_data_x, data_m = data_loader(data_name, miss_rate)

  # Impute missing data
  imputed_data_x = gain(miss_data_x, gain_parameters)
  outlier_datapoints = detect_outlier(imputed_data_x[:,0])
  #outlier_datapoints = detect_outlier(ori_data_x[:,0])
  print(outlier_datapoints)

  # Report the RMSE performance
  #rmse = rmse_loss (ori_data_x, imputed_data_x, data_m)

  print()

  return imputed_data_x

if __name__ == '__main__':

  # Inputs for the main function
  parser = argparse.ArgumentParser()
  parser.add_argument(
     '--data_name',
     #choices=['letter','spam'],
     choices=['BRFSS_pp'],
     default='BRFSS_pp',
     type=str)
  parser.add_argument(
     '--miss_rate',
     help='missing data probability',
     default=0,#originally0.2
     type=float)
  parser.add_argument(
     '--batch_size',
     help='the number of samples in mini-batch',
     default=128,
     type=int)
  parser.add_argument(
     '--hint_rate',
     help='hint probability',
     default=0.9,
     type=float)
  parser.add_argument(
     '--alpha',
     help='hyperparameter',
     default=100,
     type=float)
  parser.add_argument(
```

```
    '--iterations',
    help='number of training interations',
    default=5000,
    type=int)

  args = parser.parse_args()

  # Calls main function
  imputed_data = main(args)
```

data_loader.py

```
# Necessary packages
import numpy as np
from utils import binary_sampler
from keras.datasets import mnist

from pandas import read_csv

def data_loader (data_name, miss_rate):


  # Load data

  file_name = 'data/BRFSS_pp.csv'
  data_x = np.loadtxt(file_name, delimiter=",", skiprows=1)
  # Parameters
  no, dim = data_x.shape


  # Introduce missing data
  data_m = binary_sampler(1-miss_rate, no, dim)
  miss_data_x = data_x.copy()
  miss_data_x[data_m == 0] = np.nan

  return data_x, miss_data_x, data_m
```

gain.py

```
import tensorflow.compat.v1 as tf
tf.disable_v2_behavior()

import numpy as np
from tqdm import tqdm

from utils import normalization, renormalization, rounding
from utils import xavier_init
from utils import binary_sampler, uniform_sampler, sample_batch_index
```

```
def gain (data_x, gain_parameters):

  # Define mask matrix
  data_m = 1-np.isnan(data_x)

  # System parameters
  batch_size = gain_parameters['batch_size']
  hint_rate = gain_parameters['hint_rate']
  alpha = gain_parameters['alpha']
  iterations = gain_parameters['iterations']

  # Other parameters
  no, dim = data_x.shape

  # Hidden state dimensions
  h_dim = int(dim)

  # Normalization
  norm_data, norm_parameters = normalization(data_x)
  norm_data_x = np.nan_to_num(norm_data, 0)

  ## GAIN architecture
  # Input placeholders
  # Data vector
  X = tf.placeholder(tf.float32, shape = [None, dim])
  # Mask vector
  M = tf.placeholder(tf.float32, shape = [None, dim])
  # Hint vector
  H = tf.placeholder(tf.float32, shape = [None, dim])

  # Discriminator variables
  D_W1 = tf.Variable(xavier_init([dim*2, h_dim])) # Data + Hint as inputs
  D_b1 = tf.Variable(tf.zeros(shape = [h_dim]))

  D_W2 = tf.Variable(xavier_init([h_dim, h_dim]))
  D_b2 = tf.Variable(tf.zeros(shape = [h_dim]))

  D_W3 = tf.Variable(xavier_init([h_dim, dim]))
  D_b3 = tf.Variable(tf.zeros(shape = [dim]))  # Multi-variate outputs

  theta_D = [D_W1, D_W2, D_W3, D_b1, D_b2, D_b3]

  #Generator variables
  # Data + Mask as inputs (Random noise is in missing components)
  G_W1 = tf.Variable(xavier_init([dim*2, h_dim]))
  G_b1 = tf.Variable(tf.zeros(shape = [h_dim]))

  G_W2 = tf.Variable(xavier_init([h_dim, h_dim]))
  G_b2 = tf.Variable(tf.zeros(shape = [h_dim]))
```

```
G_W3 = tf.Variable(xavier_init([h_dim, dim]))
G_b3 = tf.Variable(tf.zeros(shape = [dim]))

theta_G = [G_W1, G_W2, G_W3, G_b1, G_b2, G_b3]

## GAIN functions
# Generator
def generator(x,m):
  # Concatenate Mask and Data
  inputs = tf.concat(values = [x, m], axis = 1)
  G_h1 = tf.nn.relu(tf.matmul(inputs, G_W1) + G_b1)
  G_h2 = tf.nn.relu(tf.matmul(G_h1, G_W2) + G_b2)
  # MinMax normalized output
  G_prob = tf.nn.sigmoid(tf.matmul(G_h2, G_W3) + G_b3)
  return G_prob

# Discriminator
def discriminator(x, h):
  # Concatenate Data and Hint
  inputs = tf.concat(values = [x, h], axis = 1)
  D_h1 = tf.nn.relu(tf.matmul(inputs, D_W1) + D_b1)
  D_h2 = tf.nn.relu(tf.matmul(D_h1, D_W2) + D_b2)
  D_logit = tf.matmul(D_h2, D_W3) + D_b3
  D_prob = tf.nn.sigmoid(D_logit)
  return D_prob

## GAIN structure
# Generator
G_sample = generator(X, M)

# Combine with observed data
Hat_X = X * M + G_sample * (1-M)

# Discriminator
D_prob = discriminator(Hat_X, H)

## GAIN loss
D_loss_temp = -tf.reduce_mean(M * tf.log(D_prob + 1e-8) \
                        + (1-M) * tf.log(1. - D_prob + 1e-8))

G_loss_temp = -tf.reduce_mean((1-M) * tf.log(D_prob + 1e-8))

MSE_loss = \
tf.reduce_mean((M * X - M * G_sample)**2) / tf.reduce_mean(M)

D_loss = D_loss_temp
G_loss = G_loss_temp + alpha * MSE_loss

## GAIN solver
D_solver = tf.train.AdamOptimizer().minimize(D_loss, var_list=theta_D)
```

```
G_solver = tf.train.AdamOptimizer().minimize(G_loss, var_list=theta_G)

## Iterations
sess = tf.Session()
sess.run(tf.global_variables_initializer())

# Start Iterations
for it in tqdm(range(iterations)):

  # Sample batch
  batch_idx = sample_batch_index(no, batch_size)
  X_mb = norm_data_x[batch_idx, :]
  M_mb = data_m[batch_idx, :]
  # Sample random vectors
  Z_mb = uniform_sampler(0, 0.01, batch_size, dim)
  # Sample hint vectors
  H_mb_temp = binary_sampler(hint_rate, batch_size, dim)
  H_mb = M_mb * H_mb_temp

  # Combine random vectors with observed vectors
  X_mb = M_mb * X_mb + (1-M_mb) * Z_mb

  _, D_loss_curr = sess.run([D_solver, D_loss_temp],
                            feed_dict = {M: M_mb, X: X_mb, H: H_mb})
  _, G_loss_curr, MSE_loss_curr = \
  sess.run([G_solver, G_loss_temp, MSE_loss],
           feed_dict = {X: X_mb, M: M_mb, H: H_mb})

## Return imputed data
Z_mb = uniform_sampler(0, 0.01, no, dim)
M_mb = data_m
X_mb = norm_data_x
X_mb = M_mb * X_mb + (1-M_mb) * Z_mb

imputed_data = sess.run([G_sample], feed_dict = {X: X_mb, M: M_mb})[0]

imputed_data = data_m * norm_data_x + (1-data_m) * imputed_data

# Renormalization
imputed_data = renormalization(imputed_data, norm_parameters)

# Rounding
imputed_data = rounding(imputed_data, data_x)

return imputed_data
```

Linear Regression Training script on BigQuery ML

```
CREATE OR REPLACE MODEL `covid_open_data.penguins_model`
OPTIONS
```

```
  (model_type='linear_reg',
    input_label_cols=['snowfall_mm']) AS
SELECT
  country_code,country_name,snowfall_mm,new_tested
FROM
  `bigquery-public-data.covid19_open_data.covid19_open_data`
WHERE
  snowfall_mm IS NOT NULL
```

Linear Regression Prediction script on BigQuery ML

```
SELECT
  *
FROM
  ML.PREDICT(MODEL `covid_open_data.penguins_model`,
    (
    SELECT
      country_code,country_name,new_tested
    FROM
      `bigquery-public-data.covid19_open_data.covid19_open_data`
    WHERE
      snowfall_mm IS NULL
      ))
```

Data Cleansing Prototype

```
from Tkinter import *
#import Tkinter.messagebox

def LinearRegression():
    from sklearn.linear_model import LinearRegression
    import time
    import numpy as np
    from sklearn import metrics
    from pandas import read_csv

    time_start = time.clock()

    train = read_csv('EHR.csv', usecols = ['Attestation_Year','Payment_Year'])
    linreg = LinearRegression()
    data = train[['Attestation_Year','Payment_Year']]
    original_DS = train.Payment_Year


    #Step-1: Split the dataset that contains the missing values and no missing values are test and
train respectively.

    x_train = data[data['Payment_Year'].notnull()].drop('Payment_Year', axis= 1)
    y_train = data[data['Payment_Year'].notnull()]['Payment_Year']
    x_test = data[data['Payment_Year'].isnull()].drop('Payment_Year', axis=1)
    y_test = data[data['Payment_Year'].isnull()]['Payment_Year']

    #Step-2: Train the machine learning algorithm

    linreg.fit(x_train, y_train)

    #Step-3: Predict the missing values in the attribute of the test data.

    predicted = linreg.predict(x_test)

    #Step-4: Let’s obtain the complete dataset by combining with the target attribute.

    train.Payment_Year[train.Payment_Year.isnull()] = predicted
    #print(train.Payment_Year) #getting the imputed data frame

    '''
    Section for applying outlier metrics for plausability evaluation
    '''

    outliers=[]
    def detect_outlier(data_1):
```

```
        threshold=3
        mean_1 = np.mean(data_1)
        std_1 =np.std(data_1)
        count = 0

        for y in data_1:
            z_score= (y - mean_1)/std_1
            if np.abs(z_score) > threshold:
                #outliers.append(y)
                count = count + 1
        return count

    # detecting outliers in original dataset
    #outlier_datapoints = detect_outlier(original_DS)
    #detecting outliers in imputed dataset
    outlier_datapoints = detect_outlier(train.Payment_Year)
    print(outlier_datapoints)
    time_elapsed = (time.clock() - time_start)
    print(time_elapsed)



def Clustering():
    from sklearn.cluster import KMeans
    from sklearn.metrics import adjusted_rand_score
    import pandas as pd
    from sklearn.feature_extraction.text import TfidfVectorizer
    import time

    time_start = time.clock()

    from pandas import read_csv

    df = read_csv('EHR.csv', usecols = ['Program_Type'])

    documents = str(df.dropna().values.tolist()).split(",")

    tfidf_vectorizer=TfidfVectorizer(use_idf=True, stop_words = 'english')

    # just send in all your docs here
    tfidf_vectorizer_vectors=tfidf_vectorizer.fit_transform(documents)
    first_vector_tfidfvectorizer=tfidf_vectorizer_vectors[0]

    # place tf-idf values in a pandas data frame
    df1 = pd.DataFrame(first_vector_tfidfvectorizer.T.todense(),
index=tfidf_vectorizer.get_feature_names(), columns=["tfidf"])

    #with pd.option_context('display.max_rows', None, 'display.max_columns', None):print(df1)
    true_k = 1
    model = KMeans(n_clusters=true_k, init='k-means++', max_iter=100, n_init=1)
```

```
    model.fit(tfidf_vectorizer_vectors)

    print("Top terms per cluster:")
    order_centroids = model.cluster_centers_.argsort()[:, ::-1]
    terms = tfidf_vectorizer.get_feature_names()
    for i in range(true_k):
        print("Cluster %d:" % i),
        for ind in order_centroids[i, :100]:
            print(' %s' % terms[ind]),
        print

    time_elapsed = (time.clock() - time_start)
    print(time_elapsed)
```

```
app = Tk()
app.title("Data Quality prototype")
app.geometry('500x500')

labelText = StringVar()
labelText.set("Proceed with missing value imputation first, then cater for noise")
label1 =  Label(app, textvariable = labelText, height = 4)
label1.pack()

button1 = Button(app, text = 'Impute with Linear Regression', width =
40,command=LinearRegression)
button1.pack()
button2 = Button(app, text = 'Noise detection', width = 40,command=Clustering)
button2.pack()
app.mainloop()
```

Appendix 3

	Accuracy	Completeness	Consistency	Availability	Validity	Usefulness	Confidence	Reliability	Provenance	Duplication
0	-0.04	0.18	0.53	0.32	0.73	0.09	-0.13	0.02	-0.31	0.27
1	0.61	0.72	0.29	0.49	0.17	0.17	0.11	0.12	-0.01	0.25
2	0.73	0.30	0.41	0.12	0.49	0.57	0.76	0.07	0.63	-0.02
3	0.74	0.77	0.60	0.59	0.65	0.35	0.57	0.23	0.27	0.38
4	0.81	0.29	0.28	0.03	0.42	0.46	0.93	-0.15	0.74	0.00
5	0.04	0.37	0.05	0.33	-0.07	-0.05	-0.07	0.25	0.31	0.12
6	0.14	0.20	0.10	0.29	-0.02	0.09	0.02	0.08	0.41	-0.09
7	0.01	-0.07	-0.14	-0.23	-0.04	0.08	-0.06	-0.01	0.47	0.32
8	0.41	0.90	0.45	0.96	0.32	0.09	0.04	0.30	-0.13	-0.01
9	0.69	0.20	0.71	0.03	0.44	0.69	0.73	0.44	0.41	0.27
10	0.40	0.12	0.30	0.03	-0.10	0.14	0.06	0.06	-0.03	0.10
11	0.45	0.28	0.09	-0.01	0.26	0.19	0.71	0.00	0.71	-0.07
12	-0.07	0.02	-0.04	-0.15	-0.29	0.10	-0.03	0.32	0.10	-0.28
13	0.03	0.09	0.43	0.03	0.02	0.48	-0.13	0.78	-0.17	-0.10
14	0.12	-0.26	0.19	-0.26	0.17	0.14	0.09	-0.21	0.12	-0.08
15	0.01	-0.03	0.19	-0.01	0.16	0.10	0.11	0.18	0.21	-0.22
16	0.05	-0.03	0.57	0.13	-0.03	0.38	-0.24	0.69	-0.27	0.00
17	0.10	0.03	0.66	0.14	0.86	0.30	0.26	0.11	0.00	0.26
18	-0.03	0.27	0.04	0.26	0.15	-0.26	-0.19	-0.01	-0.02	0.06
19	0.30	0.21	0.28	-0.04	0.21	0.46	0.32	0.45	0.34	-0.04
20	0.11	0.35	0.15	0.21	-0.20	0.22	-0.33	0.49	-0.24	-0.15
21	-0.04	-0.13	0.26	-0.16	0.02	-0.14	-0.09	0.28	0.04	0.60
22	0.18	0.58	-0.25	0.24	-0.04	-0.24	0.05	-0.03	0.21	0.07
23	0.87	0.50	0.62	0.22	0.25	0.63	0.52	0.40	0.23	0.34
24	-0.05	0.17	0.08	-0.06	0.13	-0.10	0.13	0.23	0.20	0.32
25	0.77	0.12	0.29	-0.13	0.22	0.56	0.59	-0.11	0.47	0.11
26	-0.04	0.06	0.31	0.10	0.15	-0.12	-0.15	0.30	-0.07	0.36

27	-0.11	-0.03	0.08	0.01	0.18	0.09	-0.12	0.04	0.29	0.17
28	0.05	-0.15	0.04	-0.28	-0.18	-0.14	0.12	0.07	0.28	0.12
29	0.27	0.76	0.51	0.67	0.33	0.30	-0.15	0.59	-0.32	0.18
30	0.31	0.55	0.53	0.54	0.70	0.05	0.06	-0.01	-0.13	0.41
31	0.30	0.21	0.52	0.08	0.37	0.07	0.08	0.28	0.01	0.91
32	0.05	-0.35	0.40	-0.39	0.18	0.13	0.02	0.18	0.06	0.57
33	0.25	0.04	0.56	-0.04	-0.07	0.46	0.05	0.81	0.00	0.24
34	0.23	-0.06	0.37	-0.06	0.21	0.70	0.12	0.34	0.25	-0.22
35	0.21	-0.06	-0.06	-0.13	-0.03	0.22	0.16	-0.11	0.64	0.07
36	0.07	0.18	0.24	0.42	0.05	-0.15	0.02	0.00	0.15	-0.13
37	0.02	0.05	-0.16	0.03	-0.05	0.01	-0.08	-0.12	0.48	0.05
38	0.06	-0.01	-0.06	-0.10	-0.07	0.24	-0.17	0.11	0.33	0.14
39	-0.19	-0.14	0.07	-0.05	0.44	0.00	0.02	-0.28	0.14	-0.30
40	0.57	0.39	0.76	0.41	0.20	0.38	0.08	0.44	-0.21	0.28
41	0.11	-0.05	-0.01	-0.07	-0.02	0.33	0.02	0.05	0.47	-0.04
42	0.29	0.35	0.06	0.26	0.25	0.22	0.30	0.06	0.42	-0.31

Computer Science REC
The Burroughs
Hendon
London NW4 4BT
Main Switchboard: 0208 411 5000
21/04/2021

APPLICATION NUMBER: 17925
Dear Suraj Juddoo and all collaborators/co-investigators
Re your application title: Investigating the attainment of optimum data quality for Big Data in health industry; proposing a new methodological approach

Supervisor: Carlisle George

Co-investigators/collaborators:

Thank you for submitting your application. I can confirm that your application has been given APPROVAL from the date of this letter by the Computer
Science REC.
The following documents have been reviewed and approved as part of this research ethics application:
Although your application has been approved, the reviewers of your application may have made some useful comments on your application. Please look at
your online application again to check whether the reviewers have added any comments for you to look at.
Also, please note the following:
1. Please ensure that you contact your supervisor/research ethics committee (REC) if any changes are made to the research project which could affect
your ethics approval. There is an Amendment sub-form on MORE that can be completed and submitted to your REC for further review.
2. You must notify your supervisor/REC if there is a breach in data protection management or any issues that arise that may lead to a health and safety
concern or conflict of interests.
3. If you require more time to complete your research, i.e., beyond the date specified in your application, please complete the Extension sub-form on MORE
and submit it your REC for review.
4. Please quote the application number in any correspondence.
5. It is important that you retain this document as evidence of research ethics approval, as it may be required for submission to external bodies (e.g., NHS,
grant awarding bodies) or as part of your research report, dissemination (e.g., journal articles) and data management plan.
6. Also, please forward any other information that would be helpful in enhancing our application form and procedures - please contact
MOREsupport@mdx.ac.uk to provide feedback.
Good luck with your research.

Yours sincerely
Gill Whitney

Member Computer Science REC