Middlesex University

VAST 2010 Challenge
Hospitalization Records - Characterization of Pandemic Spread

Authors and Affiliations:

Peter Passmore, School of Engineering and Information Sciences, Middlesex University, UK, p.passmore@mdx.ac.uk
Yongjun Zheng, School of Engineering and Information Sciences, Middlesex University, UK, y.zheng@mdx.ac.uk
Chris Rooney, School of Engineering and Information Sciences, Middlesex University, UK, c.rooney@mdx.ac.uk
Tamara Al-Sheikh, School of Engineering and Information Sciences, Middlesex University, UK, t.al-sheikh@mdx.ac.uk
Kai Xu, School of Engineering and Information Sciences, Middlesex University, UK, k.xu@mdx.ac.uk [PRIMARY contact]

Tool(s):

Microsoft Excel 2007
KNIME http://www.knime.org/
Java
JFreeChart: http://www.jfree.org/jfreechart/
C++

Video:

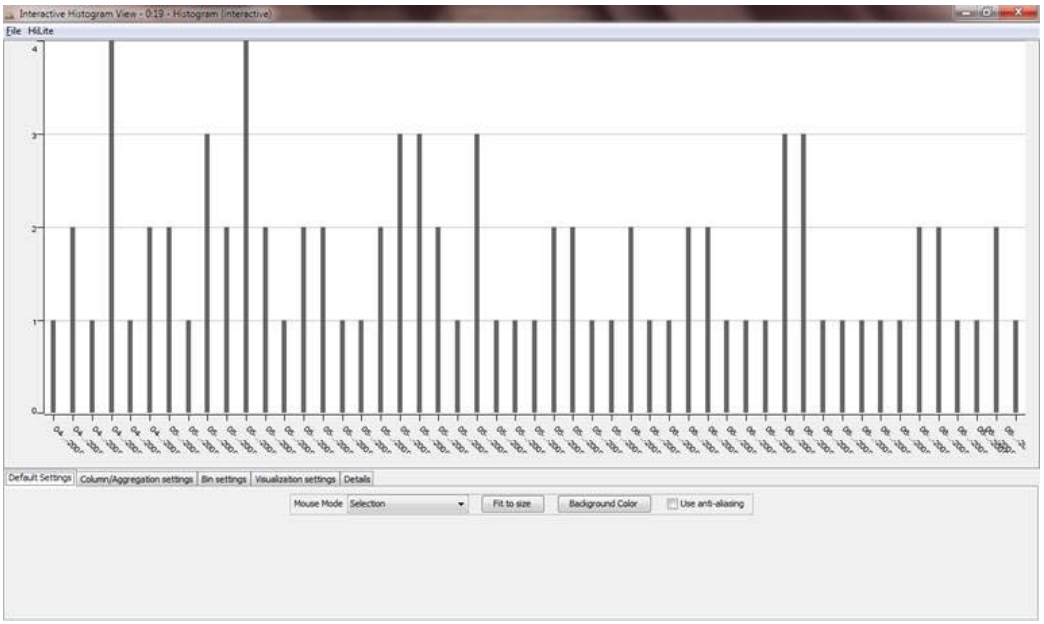
Video

ANSWERS:

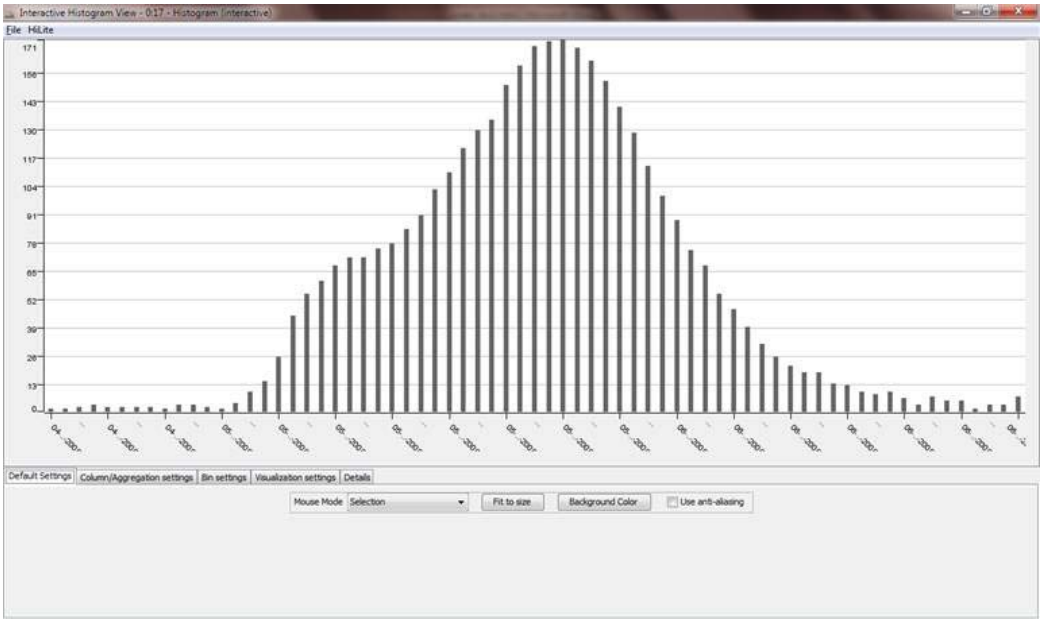
MC2.1: Analyze the records you have been given to characterize the spread of the disease. You should take into consideration symptoms of the disease, mortality rates, temporal patterns of the onset, peak and recovery of the disease. Health officials hope that whatever tools are developed to analyze this data might be available for the next epidemic outbreak. They are looking for visualization tools that will save them analysis time so they can react quickly.

Patient death frequency

We started by checking the number of patient death over time. We used the KNIME initially for a quick check, because it requires little or no programming. We plotted number of death against time for each country. The result of Thailand didn't show anything interesting (see the figure below)



Whereas the result of Venezuela (figure below) clears show a peak in patient death number.



We then use Java and JFreeChart to do the same plot for all other countries. The result shows that all countries except Thailand and Turkey have a peak in death. We suspect the peak in patient death is related to epidemic.

Syndrome

Using Microsoft Excel 2007, we found there are about 1200 distinct strings in SYNDROME column. However, by manual inspection, we found many of strings are describing the same syndrome, such as AB PAIN, ABD PAIN, and ABD.PAIN.

We counted the frequency of people dying with each symptom by joining the patient record and death record. We generated spreadsheet that shows the results side by side. Quick visual analysis of numbers shows very low numbers for Thailand and Turkey. This confirmed our previous conjecture and we discarded these two countries as having no sign of epidemic.

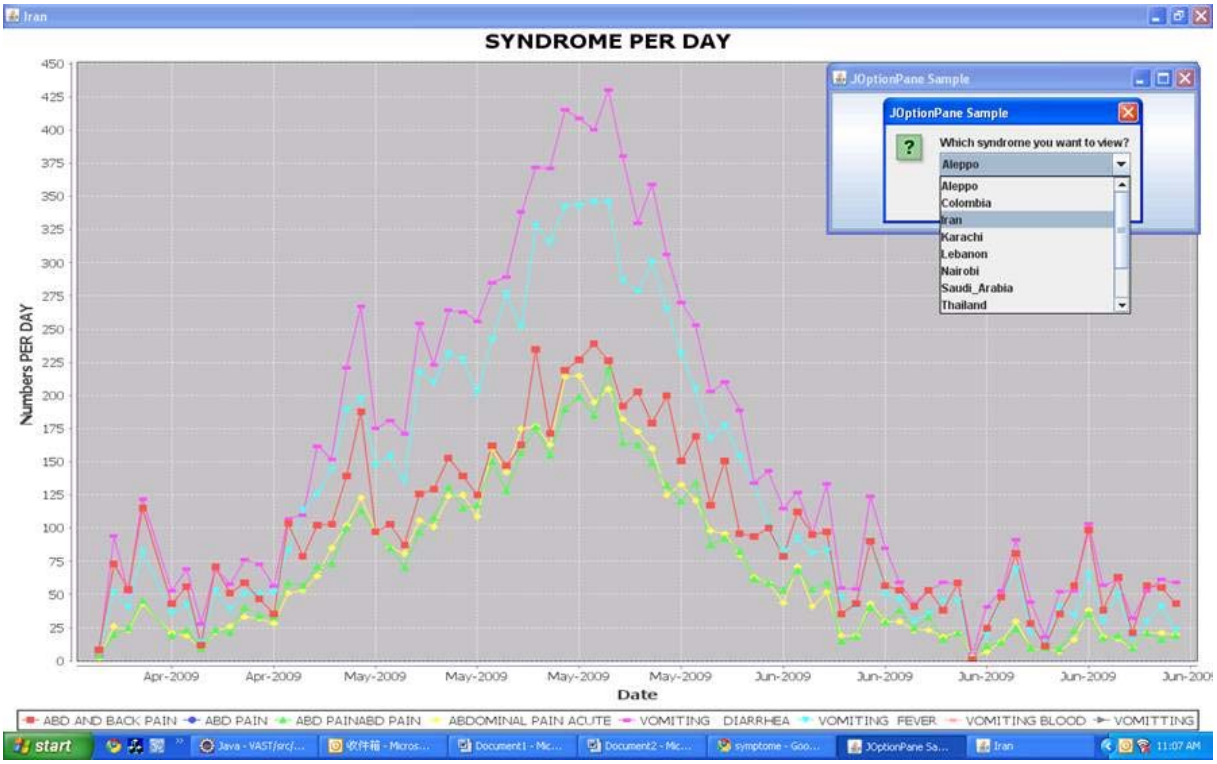
For all 9 remaining countries, we found a sudden falloff in numbers from position 75 to 76 after order symptoms by number of deaths. The top 75 symptoms account for 94% to 97% of numbers and they are the same for each country. Therefore we decided to focus on top 75 symptoms listed below:

1	ABD AND BACK PAIN	26	BACK INJ	51	TREMORS
2	ABD CRAMPING	27	BACK INJURY	52	VOMITING
3	ABD PAIN	28	BACK PAIN	53	VOMITING DIARRHEA
4	ABD PAIN FEVER	29	BACK PAIN STRAIN	54	VOMITING BLURRED VISION
5	ABD PAIN VOMITING	30	BACK PAINBACK PAIN	55	VOMITING DIARRHEA
6	ABD PAINABD PAIN	31	BACK PAINS	56	VOMITING FEVER
7	ABD PAINFEVER	32	BACK PAINSTRAIN	57	VOMITING & DIARRHEA
8	ABD PAINS	33	BACK PN	58	VOMITING ABD PAIN
9	ABD PAINVOMITING	34	BACK PX	59	VOMITING ALL DAY
10	ABD PN	35	BACK SPASMS	60	VOMITING ALONE
11	ABD PX	36	BACK STRAIN	61	VOMITING AND DIARRHEA
12	ABD. PAIN	37	CONJUNCTIVITIS RED	62	VOMITING AND FEVER
13	ABD.PAIN	38	DIARRHEA	63	VOMITING BLOOD
14	ABDBACK PAIN	39	DIARRHEA AND VOMITING	64	VOMITING BLOODVOMITING BLOOD
15	ABDMNAL PAIN OTH SPCF ST	40	DIARRHEA FEVER	65	VOMITING DIARRHEA
16	ABDMNAL PAIN UNSPCF SITE	41	DIARRHEA VOMITING	66	VOMITING DIARRHEA FEVER
17	ABDOMEN PAIN	42	ENCEPHALITIS	67	VOMITING FEVER
18	ABDOMINAL CRAMPING	43	FACIAL SWELLING	68	VOMITING RASH
19	ABDOMINAL PAIN	44	HEAD	69	VOMITINGABD PAIN
20	ABDOMINAL PAIN ACUTE	45	HEARING LOSS	70	VOMITINGDIARRHEA
21	ABDOMINAL PAIN VOMITING	46	NOSE	71	VOMITINGFEVER
22	ABDOMINAL PAINABDOMINAL PAIN	47	NOSE BLEED	72	VOMITINGHEADACHE
23	ABNORMAL LABS	48	NOSE BLEEDNOSE BLEED	73	VOMITTING
24	BACK AND LEG PAIN	49	NOSE BLEEDS	74	VOMITTING BLOOD
25	BACK AND NECK PAIN	50	PROTEINURIA	75	VOMTING

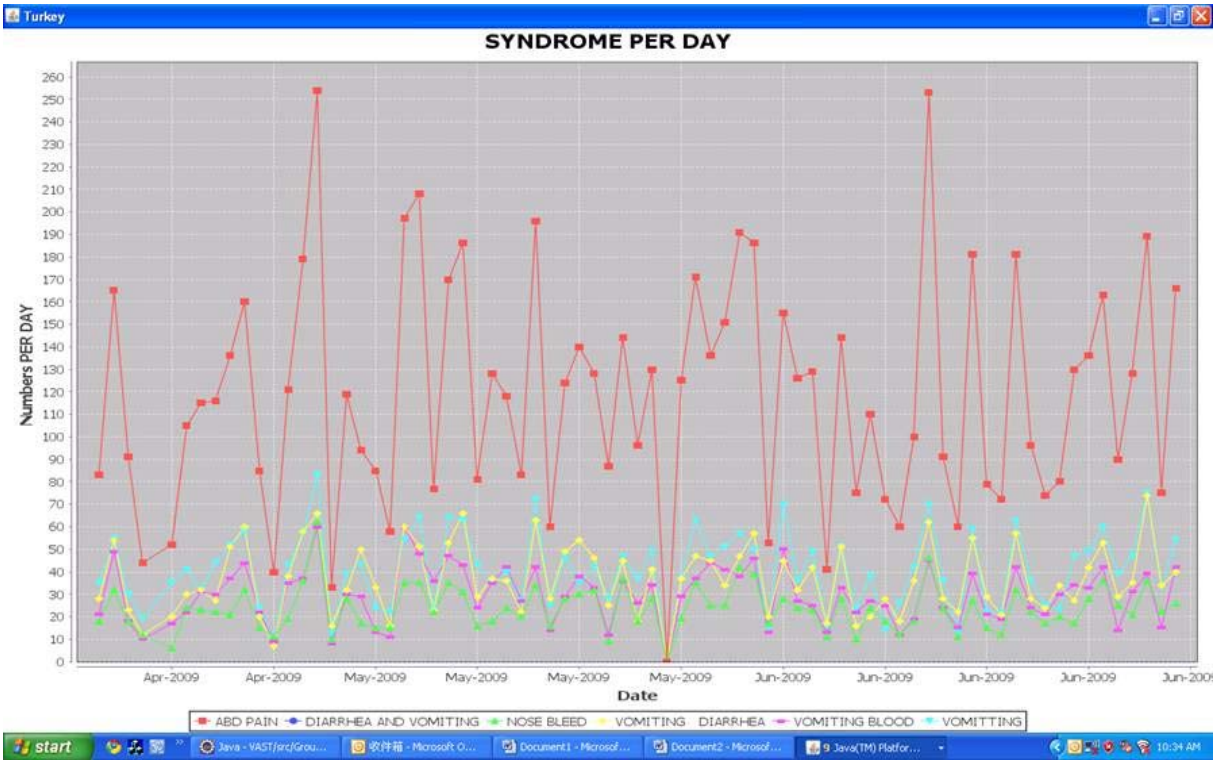
We then categorize these symptoms by grouping any symptoms that contains “vomit” as VOMITING, “abd” as ABDOMINAL PAIN, etc. We then ordered them according to frequency and found there is a considerable drop between 5th and 6th and the top 5 symptoms are: VOMTING, ABDOMINAL PAIN, BACK, DIARRHEA, and NOSE BLEED.

We also looked at the similarity of symptom frequency change in different countries. We assumed that the symptoms of the epidemic will have curves over time, whereas unrelated symptoms will have quite different curves. By finding the most similar curves, we can identify the symptoms associated with the epidemic. We computed the pair-wise Cosine Similarity between all symptom curves and select the top group of curves for all countries. The implementation is done in Java and the most similar curves are plotted using JFreeChart.

The results confirmed our previous findings (see the figure below): the identified symptoms matched well with the top 5 discussed before (these are listed in the bottom of the figure below and they are not combined); also the curve is very similar to that of the death frequency. A simple interface is implemented so we can select a country from a drop-down list and the computation is then done for that country and results displayed.

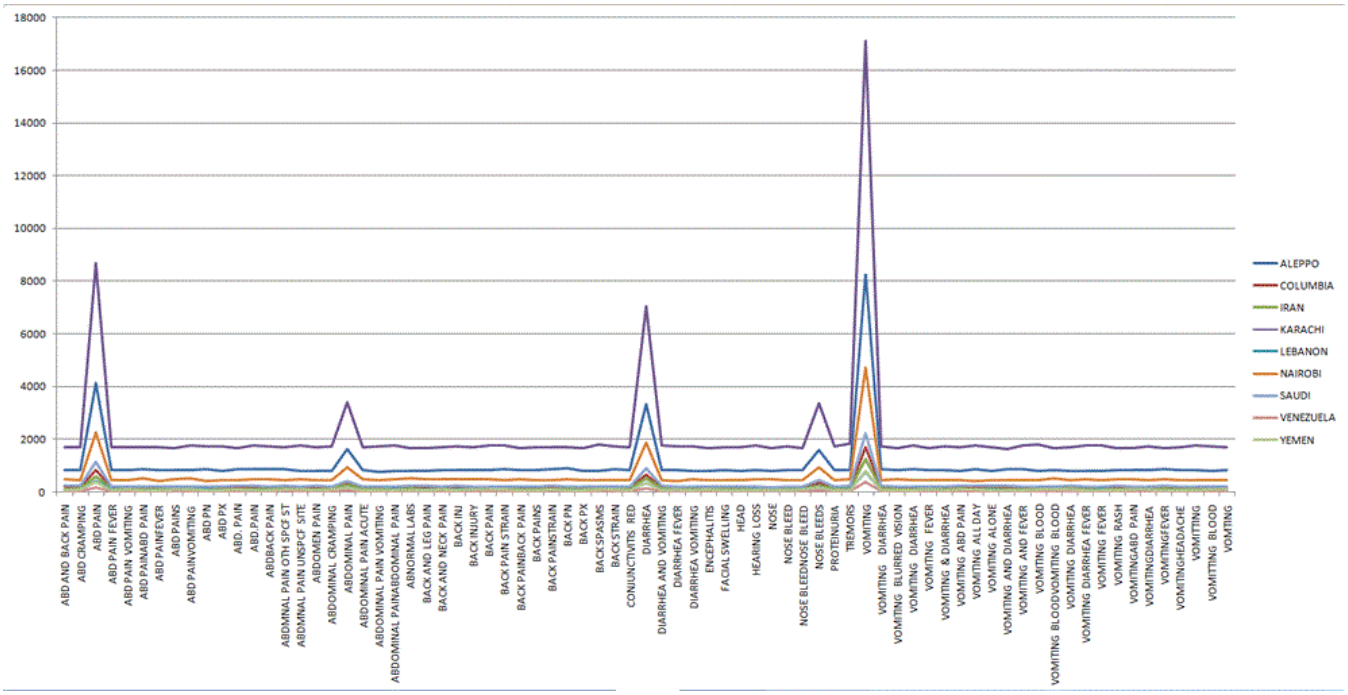


Again, the result of Thailand and Turkey did not show any overall trend (the plot of Turkey is shown below).

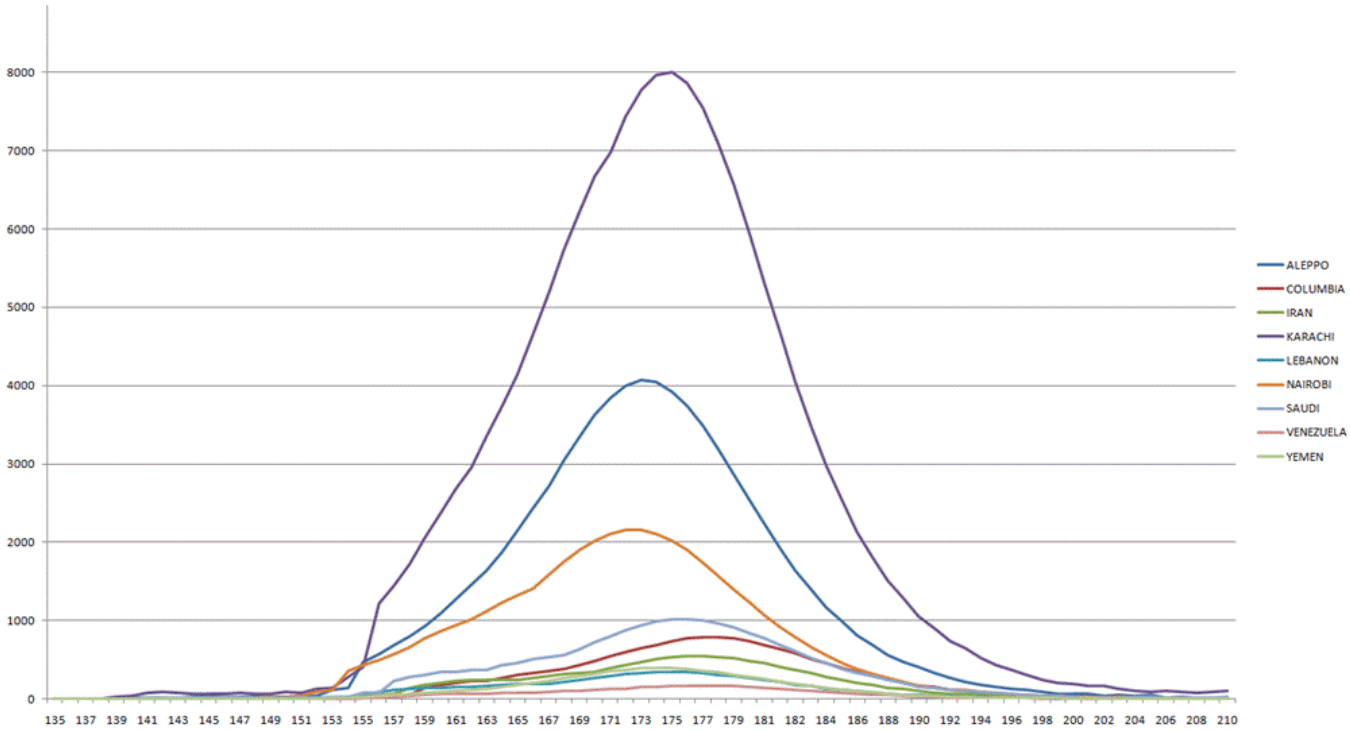


MC2.2: Compare the outbreak across cities. Factors to consider include timing of outbreaks, numbers of people infected and recovery ability of the individual cities. Identify any anomalies you found.

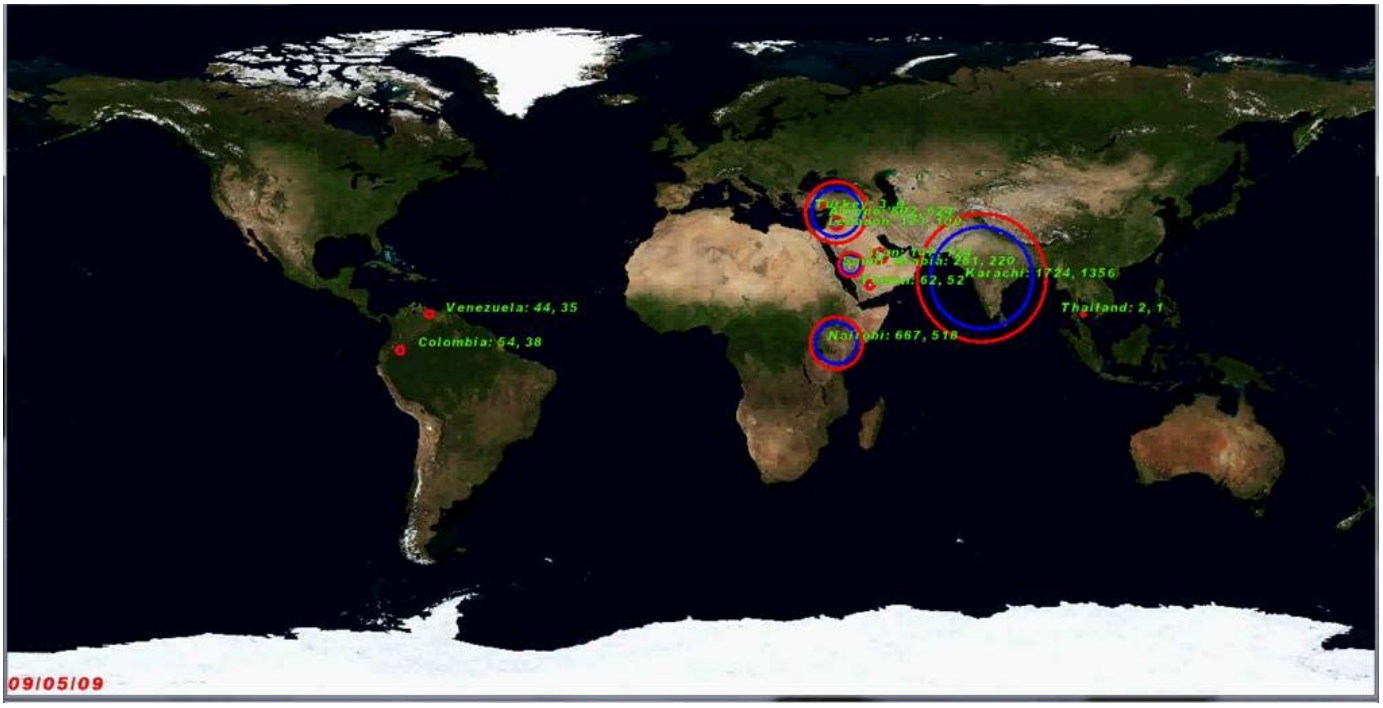
The graph below shows the number of patient of the top 75 symptoms for the 9 countries with epidemic (produced with Microsoft Excel and so are the rest). The peaks match with our previous analysis.



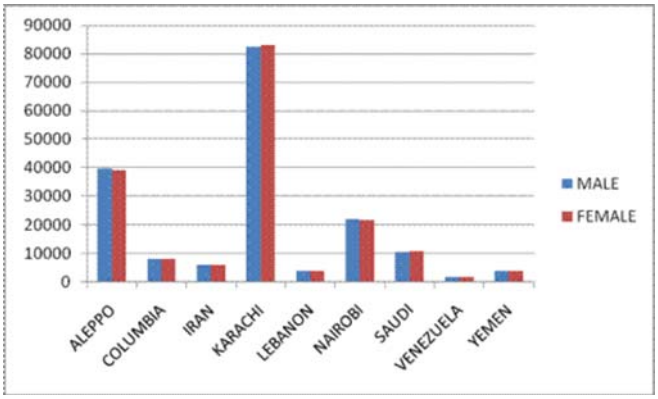
The graph below shows the daily death count by location. Karachi, Aleppo and Nairobi seem to start earlier and are more severe than the others.



We also produced an animation to show number of patient death over time in the 9 countries with epidemic (below is a screenshot). The red circle represents the total number of deaths, the blue circle represents the number of deaths from the top 75 symptoms. The text represents the country, total deaths, and deaths from the 75 symptoms. The date can be seen in the bottom left corner.



We also checked the distribution by sex, but did not find any significant effect.



The graph below shows the distribution by age category. All the countries have a similar pattern: the more severe the epidemic the more people die in the middle age range.