18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Characteristics of outbreak-response databases: Canadian SARS example

Cuff, Wilfred R. [1], Binhua Liang [2] Venkata R. Duvvuri [3] and Jianhong Wu [3,4]

[1]Public Health Agency of Canada, Canadian Science Centre for Human and Animal Health, Winnipeg, Manitoba, Canada; [2]Medical Microbiology, University of Manitoba, Winnipeg, Manitoba, Canada; [3]Centre for Disease Modelling, York University, Toronto, Ontario, Canada; [4]Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada

Email: Wilfred_Cuff@phac-aspc.gc.ca

Abstract: In 2003, SARS was a serious health concern in Canada. As of September 3 of that year, the Public Health Agency of Canada reported a total of 438 cases: 251 Probable (247 Ontario, 4 British Columbia) and 187 Suspect (128 Ontario, 46 British Columbia). Although the outbreak was short-lived, more than forty people died from the disease.

A substantial database evolved as a consequence of control efforts. Specimens began to be received and tested at the National Microbiology Laboratory (NML) in Winnipeg, Manitoba, on March 17, 2003. NML's SARS database contains more than 12,000 records and 192 variables, with variables detailing clinical/ diagnostic (17), microbiological (143), epidemiological (25), and administrative (7) features. Clinical variables include: diarrhea, difficulty of breathing, severity of illness, systemic status, date of onset of illness, case status, case status modification, and case status modification date. Diagnostic variables include: fever, chest X-ray change, cough, shortness of breath, contact with probable case, travel, source of exposure, and contact type. Epidemiological data include: date of birth, age, sex, epidemiology cluster, and employment status. Administrative data include: patient's last name, first name, temporal data (date of collection of the specimen, date of its receipt, and first date of hospitalization of the patient) and spatial data (origin of specimen, and identity of the hospital).

A wide variety of laboratory tests are included in the database: Enzyme-Linked Immunosorbent Assay (ELISA, 7 variables), Immunofluorescence Assay (IFA, 7); plaque reduction neutralization test (PRN, 3); cytopathogenic test (CPE, 2), and electron microscopy test (EM, 8). There are tests for Coronavirus (13 variables), human metapneumovirus (hMPV, 16), circovirus (Circo, 4), porcine circovirus (PCV1, 6), TT-virus (TTV, 7), TTV-like-mini-virus (TLMV, 7), Hantaanvirus (1), Rhinovirus (3), and Paramyxovirus (3). Nested PCR, RT-PCR, and sequencing tests are common among the viruses.

The NML-SARS database evolved as part of an ongoing effort involving multiple institutions, multiple regions, intense time pressures, and the participation of many operational and scientific specialties (*e.g.,* clinicians, epidemiologists, microbiologists, administrators). Although the database arose in response to a specific disease in Canada, it can be looked upon as an example of what might typically arise from a public-health response to an outbreak of an emerging disease. Hence there is value in analysing the NML-SARS database, looking for general characteristics, and highlighting where opportunities for scientific advances exist. This is our objective.

Putative characteristics of outbreak-response data sets are: *ad hoc* by definition; evolving database and data administration; basic assumptions (*e.g.,* case definition) open to refinement; and insights that are often merely suggestive. Opportunities for scientific advancement include: Exploratory Data Analysis of evolving data sets; definition of a relational data model appropriate to outbreak-response data sets; improvement of statistical data modelling methodology to estimate empty blocks of cells (resulting as emerging understanding directs interest from one area to another); data analysis to refine basic assumptions made during control operations; process modelling to explore consequences of unverified insights.

Keywords: *SARS, outbreak-response data sets, relational data model, disease modelling*

1. INTRODUCTION

SARS was a serious health concern in Canada in 2003. Specimens began to be received and tested at the National Microbiology Laboratory (NML) in Winnipeg, Manitoba, on March 17, 2003. As of August 5, there were a total of 438 cases in Canada: 251 Probable and 187 Suspect. The outbreak was short-lived but more than forty Canadians died from the disease.

The Canadian SARS operations were co-ordinated by a state-of-the-art facility situated in Winnipeg, Manitoba. The Canadian Science Centre for Human and Animal Health is an agency of the Public Health Agency of Canada, and houses the National Microbiology Laboratory (NML).

2. THE NML-SARS DATABASE

2.1. Overview

Laboratory results were linked with clinical, diagnostic, epidemiological, and administrative data from different provinces and hospitals to generate a substantial data set.

2.2. Variables

There are 192 variables in the database, roughly split into classes as shown in Table 1. Microbiological (or laboratory-based) data constitute most of the variables (about 75%).

This group can be classified into viral-specific and non-specific variables. In the viral-specific class, there are tests for Coronavirus (13 variables), human metapneumovirus (hMPV, 16), circovirus (Circo, 4), porcine circovirus (PCV1, 6), TT-virus (TTV, 7), TTV-like-mini-virus (TLMV, 7), Hantaanvirus (1), Rhinovirus (3), and Paramyxovirus (3). For each virus, variables describe results from various tests or detail laboratory administration. Nested PCR, RT-PCR, and sequencing tests are common, as are administrative variables giving the name of lab investigator or name of primer.

Table 1. Number of Variables by Class

Class	Variables
Clinical/Diagnostic	17
Microbiological	143
Epidemiological	25
Administrative	7

Only 9 of the Microbiological variables containing test data are populated with more than 100 rows of data; and only 4 of these with more than 1000 rows. Over time, many pathogens and tests were evaluated and discarded as they produced negative findings. This resulted in virtual subsets of data related to specific tests and/or organisms.

In the non-specific laboratory class, the variables are grouped simply by type of test: Enzyme-Linked Immunosorbent Assay (ELISA, 7 variables), Immunofluorescence Assay (IFA, 7); plaque reduction neutralization test (PRN, 3); cytopathogenic test (CPE, 2), and electron microscopy test (EM, 8).

Epidemiological data include variables such as date of birth, age, sex, epidemiology cluster, and employment status.

Clinical variables include signs and symptoms such as diarrhea, difficulty of breathing, and severity of illness; and additional related variables such as systemic status, date of onset of illness, case status, case status modification, and case status modification date. Diagnostic variables include: fever, chest X-ray change, cough, shortness of breath, contact with probable case, travel, source of exposure, and contact type.

Administrative data include variables such as: patient's last name, first name, temporal data (date of collection of the specimen, date of its receipt, and first date of hospitalization of the patient) and spatial data (origin of specimen, and identity of the hospital).

2.3. Data Storage

As a consequence of the emergency, the database had to be set up quickly, using *ad hoc* staff. The database model had to be defined at a time when there was minimal understanding of the characteristics of this emerging disease, and had to be capable of modification as knowledge accumulated. The software needed to be capable of reliably generating simple, routine reports.

The NML-SARS data were stored simply as a large flat file, in a "user friendly" relational database management system (rdbms). The data are presented to the database manager by the software as one large spreadsheet and to users as Views, set up to serve specific purposes.

Eighty-five columns (out of a total of 237) in the NML-SARS dbms are "global columns" - containing only one value - and used in "summary" and "calculation" columns, which are derived columns used for operational reporting. Derived columns are presented in the same way as data columns. There are also many "comment" columns in the database.

2.4. Database Documentation & Maintenance

A 4-page table (Glossary of Database Fields) documents the NML-SARS data, with each column described by at most a few lines. The levels of categorical columns are not always defined. Recourse to relevant specialists for complete understanding is necessary.

The NML-SARS data set and database were generated to serve control operations during the outbreak.

2.5. Opportunity: Relational Data Model

It seems to us that outbreak-response data could benefit from being organized by a relational data model. In the flat-file NML-SARS database it is unclear whether the data were organized by patient, by measurement on patient, or by laboratory test. Of the 14,234 records in the database, there was no unique record identifier: Patient Number contained 12,693 entries, the NML Number 14,185, and the Unique Specimen Number 14,227.

The large and diverse set of columns made the data difficult to extract with confidence. It proved most effective to extract simple subsets of data related to specific questions. Even verification of data quality within the database was difficult.

The diversity of multiple records per patient (in the absence of a relational model) was also a major impediment in obtaining a data subset appropriate for data analysis.

Fortunately, a relational model for a general outbreak-response data set can be envisaged. A Master Table providing patient data seems sensible, and would include a key column/s and pointers to various other tables like Patient_Administration, Diagnostic, Clinical, Hospital, Laboratory_Test, Geo-referencing, *etc*.

As an example of what such a relational model might look like, we provide in Figure 1 the main part of the laboratory component of an outbreak-response data set.

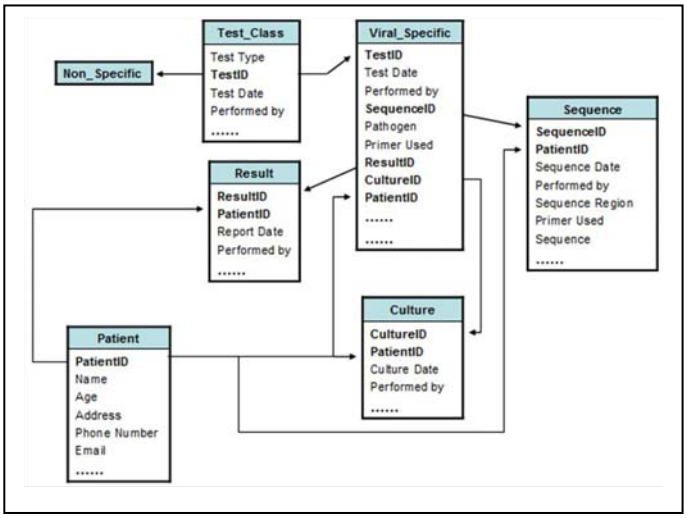

Figure 1. Proposed data model for laboratory-based part of an outbreak-response data set.

The Viral_specific table is fully specified (for the NML-SAR case) and Non-specific could similarly be elaborated.

3. DATA ANALYSIS

As the largest percentage of the columns in the database concerns laboratory tests (Table 1), this subset was analysed to infer general, statistical properties of outbreak-response data sets.

3.1. Much is Open to Refinement

Laboratory variables that are represented by more than 1000 rows are: ELISA Result 1, IFA Result 1, Coronavirus Nested PCR, Coronavirus RT-PCR, and a large number of variables concerned with laboratory-administration. Variables represented by more than 100 but less than 1000 rows are: PRN, Titer PRN,

hMPV Nested PCR, PCR hMPV, RT-PCR hMPV, Sequence Coronavirus (92), and some variables concerned with laboratory-administration. Emphasis will be placed on these columns.

Anderson *et al.* (2005) noted that as SARS was a new disease during the data-collection period, the case definition should be interpreted with caution. Accordingly, each the two measures of the incidence of SARS (Severity of Illness, Case Status) were related to the variables identified in the WHO case definition, *vis.,* fever, chest X-ray change, cough, shortness of breath, contact with a probable case, and travel (Riley *et al.*, 2003). (Case Status has values of Control, Neither, Probable, Suspect, Under Investigation and Severity of Illness has values of Deceased, Respirat, blank).

A binary logistic model was fit to each incidence measure. Case Status was made binomial, as Probable and the aggregate of Control, Neither, Suspect, and Under Investigation. The 11 rows in which Severity of Illness had the value Respirat related to only one patient and were eliminated. The model fits suggested the relative unimportance of cough, shortness of breath, and travel (Case Status only).

Furthermore, as shown in Table 2, Severity of Illness provides better discrimination than Case Status.

Table 2. Predictive ability of a binary logistic model, using two measures of incidence of SARS. Top table: Case Definition; bottom table: Severity of Illness

		Predicted	
		Other	Probable
Observed	Other	71	5
	Probable	11	41 (17 patients)

		Predicted	
		Alive	Deceased
Observed	Alive	285	6
	Deceased	4	23 (3 patients)

3.2. Much is Suggestive, not Verified

A major question addressed in the control operations was: what is the causative agent of SARS?

Virus	Test	Positive Records	Negative Records
	PRN	63	72
	ELISA Reslt 1	361	5273
	IFA Reslt 1	367	5278
Coronavirus	Nested PCR	475	7793
Coronavirus	RT-PCR	36	1045
Coronavirus	Sequence	87	5
hMPV	Nested PCR	18	202
hMPV	RT-PCR	17	468
hMPV	PCR	1	271
TT-virus	PCR	4	3
TTV-like mini-virus	PCR	4	3

Table 3. Test details for the most populated Laboratory measurements

Poutanen *et al.* (2003) noted, on "the basis of preliminary investigations, it appears that this syndrome may be due in part to the newly described respiratory viral pathogen, human metapneumovirus, to a novel coronavirus, or both". The NML-SARS database was used to explore this issue.

Sample size of the most populated laboratory measures is shown in Table 3; all variables are categorical and sample sizes are shown by value. Sample size for data analysis of pathogen is adequate only for Coronavirus and hMPV (not including PCR). Simple cross-tabulations suggested a strong relationship between ELISA Result 1 and IFA Result 1; a small but consistent inverse relationship between each of ELISA Result 1 and IFA Result 1, and each of Nested PCR and of RT-PCR (IFA only) – for both Coronavirus and hMPV; an inverse relationship between Nested PCR and RT-PCR for both Coronavirus and hMPV (but sample size was small); and a direct relationship between Nested PCR and Sequence for Coronavirus.

Based upon these correlations, a binary logistic model was run (full factorial, forward stepwise model, n=14,211 records) to relate Severity of Illness to PRN, ELISA Result 1, hMPV Nested PCR, Coronavirus Nested PCR. Good

discrimination was found for Alive but for Deceased it was only ~50%. (R^2 was only 17%.) The largest coefficients were ELISA Result 1 (decreases probability of death) and PRN (increases it). The only significant coefficient was hMPV Nested PCR (decreases probability of death). Coronavirus Nested PCR entered the model only as an interaction between Coronavirus and hMPV.

When RT-PCR for both Coronavirus and hMPV were added to the model, the fit rose to an 80% correct assessment of Deceased. (R^2 rose to 24%.) Notably ELISA Result 1 did not enter the model as factor. Coronavirus Nested PCR now entered the model as factor (increases probability of death significantly, OR=~4). RT-PCR Coronavirus and hMPV both decrease probability death, with a larger parameter value for hMPV.

These analyses support the assertion of Poutanen *et al.* (2003) that both Coronavirus and hMPV may be involved in SARS, but the data are inadequate to make a definitive conclusion.

3.3. Opportunities

The fact that the mining of outbreak-response data may be inadequate to provide statistical significance suggests modelling opportunities of two sorts.

3.3.1. Impute Missing Blocks of Data

There is a need for imputing missing values. The challenge in outbreak-response data sets is different from conventional statistical approaches in that typically blocks of data - rather than individual values - are "missing".

This happens because events change as the database is being filled; understanding evolves. The putative pathogens change as some are ruled out and new ones begin to be investigated. The diagnostic measurements which clinicians/hospitals supply change over time. The organizations contributing data change over time, and their variables with them.

These changes are obviously not desirable from the viewpoint of data analysis, as is clear from the previous section. Any progress leading to stronger inferences would be useful. Serious problems with missing values are also characteristic of microarray data sets and new approaches are arising in response (*e.g.,* Wang *et al.*, 2006) that might be helpful here.

3.3.2. Model to Infer Consequences of Weak Inferences

Another opportunity is for process modellers to explore consequences of weak inferences raised from analysis of the data.

Two models of particular relevance to Canada were published by Choi & Pak (2003) and by Gumel *et al.* (2004). The statistical modelling reported earlier concerns the log odds of an observation being classified as Deceased, as a result of the presence/absence of infection by Coronavirus and/or hMPV. The presence of multiple pathogens suggests changes to existing models in any of a number of ways, and such improvements could potentially lead to novel findings related to non-trivial transmission patterns. Taking the model of Gumel *et al.* (2004) as an example, it is clear that any of the transmission coefficient β; rate of development of clinical symptoms $\kappa 1$ & $\kappa 2$; rate of disease-induced death d1 & d2; and rate of disease-induced recovery $\sigma 1$ & $\sigma 2$ could be affected by the presence of the two putative pathogens and the nature of their effect.

Modellers have noted the existence of super-spreading events, those "rare events where, in a particular setting, an individual may generate many more than the average number of secondary cases" (Riley *et al.*, 2003). These authors speculate that it "may be that the distinction between typical infection events and SSEs ["super-spread events"] reflects ... different routes of transmission", *i.e.*, respiratory exudates and fecal-oral contact. That may be so, but there was an early reference in the literature (Poutanen *et al.,* 2003) and data-based reasons (analyses reported here, and the recent publication of Lee *et al.,* 2007) for expecting the existence of two, possibly antagonistic, pathogens involved in SARS. Riley *et al.* (2003) give a start ($R_t = R_t^{XSS} + p^{SSE}N^{SSE}$), where SSE represents a Super-Spreading Event; XSS is a "normal-spread" Event; R_t is the population Reproduction Number at time t; R_t^{XSS} is that Number for the proportion of the population that is a XSS; p^{SSE} is the probability of an SSE; and N^{SSE} is the number of individuals participating in an SSE.

It may prove useful to partition the Infectives in terms of the pathogen, within the hosts and their status of co-infection. Parameterizing the model may induce a re-visit of the outbreak-response data set in order to examine the correlation between co-infection and super-spread events.

Gumel *et al.* (2004) deal extensively with an "optimal isolation program" and it would be useful to explore the consequences of an assumption of two antagonistic pathogens on their recommendations. A particular issue of great Canadian interest is whether the observed twin peaks of SARS outbreak in several cities were truly random events, or a deterministic outcome of the co-existence of two pathogens.

4. DISCUSSION AND CONCLUSIONS

A number of characteristics of outbreak-response data sets have been proposed following from an investigation of the Canadian NML-SARS database. This database emerged from a public-health response to an unexpected outbreak of a disease with serious implications. It evolved as part of an effort involving multiple institutions, multiple regions, intense time pressures, and involving many operational and scientific specialities. The database had to be set up quickly.

As the database model had to be defined at a time when there was minimal understanding of the characteristics of this emerging disease, it evolved. Likewise, project and database administration were brought together specifically to deal with the emergency. Hence, aspects of project management, database administration, and scientific and technical contributions also evolved over time. Fundamental decisions and guidelines (like case definition) were made as needed, and also evolved over time.

The resulting data set and database incorporate features that are not normally found in data collected and organized for purely scientific purposes. The database includes many classes of variables (microbiological, epidemiological, clinical, administrative), many variables (237), many observations (14,234), many house-keeping variables, and hence many blocks of missing values.

In consequence, it was difficult to understand the order in the data and hence to make reliable exports for analysis. This aside, it was still hard to draw reliable inferences.

But these challenges represent opportunities. A rational Relational Data Model for outbreak-response data sets would be immediately useful. Exploratory Data Analysis of diverse and sparse data sets needs some innovative work, including further work on statistical data modelling to estimate empty cells and (especially) blocks of cells. Statisticians with access to the data can make important contributions in real time simply by evaluating and refining the basic assumptions used in the control operations. Since firm conclusions may be hard to come by, so there is an opportunity for modellers to explore consequences of possibly important ideas, weak inferences suggested by the data.

Thus, three types of modelling opportunities are thus presented by outbreak-response data sets. First, there is a need for an appropriate data model for a RDBMS. Secondly, the very sparse data set that results as emerging understanding directs interest from one area to another severely limits the range of statistical investigations possible; and this provides the opportunity to estimate blocks of missing values by data modelling. Lastly, process modelling is called for to explore consequences of ideas that are suggested by ongoing investigations.

5. ACKNOWLEDGMENTS

WRC thanks the Public Health Agency for access to the NML-SARS database, and support in conducting these analyses. VRD and JW would like to acknowledge support from the Canadian Network of Centers of Excellence MITACS.

6. REFERENCES

Anderson, R.M., Fraser, C., Ghani, A.C., Donnelly, C.A., Riley, S., Ferguson, N.M., Leung, G.M., Lam, T.H., and Hedley, A.J. (2005), Epidemiology, transmission dynamics, and control of SARS: the 2002-2003 epidemic. Chapter 10 *in*: Eds: McLean, A.R., May, R.M., Pattison, J., and Weiss, R.A., *SARS A Case Study in Emerging Infections,* Oxford University Press. 61-80.

Choi. B. C. K. and Pak, A. W. P. (2003), A simple approximate mathematical model to predict the number of severe acute respiratory syndrome cases and deaths. *J. Epidemiol Community Health,* 57, 831-835.

Gumel, A. B., Ruan, S., Day, T., Watmough, J., Brauer, F., van den Driessche, P. Gabrielson D, Bowman C, Alexander ME, Ardal S, Wu J and Sahai BM (2004), Modelling strategies for controlling SARS outbreaks. Proc. R. Soc. Lond. B. doi:10.1098/rspb.2004.2800.

Lee, N., Chan, P. K., Yu, I. T., Tsoi, K. K., Lui, G., Sung, J. J., and Cockram, C. S. (2007), Co-circulation of human metapneumovirus and SARS-associated coronavirus during a major nosocomial SARS outbreak in Hong Kong. *Journal of Clinical Virology* 40, 333-337.

Poutanen SM, Low DE, Henry B, Finkelstein S, Rose D, Green K, Tellier R, Draker R, Adachi D, Ayers M, Chan AK, Skowronski DM, Salit I, Simor AE, Slutsky AS, Doyle PW, Krajden M, Petric M, Brunham RC, McGeer AJ (2003), Identification of severe acute respiratory syndrome in Canada. *New Engl. J. Med.* 348, 1995-2005.

Riley, S., Fraser, C., Donnelly, C.A., Ghani, A.C., Abu-Raddad, L.J., Hedley A.J., Leung GM, Ho LM, Lam TH, Thach TQ, Chau P, Chan KP, Lo SV, Leung PY, Tsang T, Ho W, Lee KH, Lau EM, Ferguson NM, Anderson RM. (2003), Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* 300, 1961-66.

Wang, X., Li, A., Jiang, Z., and Feng, H. (2006), Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics* 7, 32.