2011

Some Methodological Issues in Biosurveillance

Fricker, Ronald D. Jr.

Fricker, R.D., Jr. (2011). Some Methodological Issues in Biosurveillance (with commentaries [1] [2] [3] [4] [5] and rejoinder), Statistics in Medicine, 30, 403-441.
https://hdl.handle.net/10945/38755

Commentary

Statistics in Medicine

(wileyonlinelibrary.com) DOI: 10.1002/sim.4112

Published online in Wiley Online Library

Comments on 'Some methodological issues in biosurveillance'‡

Henry R. Rolka*†

I would first like to compliment Dr Fricker on the outreaching, cross-disciplinary nature of this work. Underlying his discussion of several important methodological issues in biosurveillance is an important theme: successful methods and approaches from other fields of endeavor have much to offer for the analysis and exploitation of public health data. Dr Fricker focuses on three topics from which applied public health surveillance and biosurveillance can certainly benefit: techniques for measuring change from the field of statistical process control, examples from perception and cognition theory (e.g. the 'looking for everything' analogy to sydromic surveillance), and the value of using simulation for substantiating new methodological applications. His development of concepts remains rooted in the goals of early disease event detection in a population and extends to the less rigorously defined but more intuitive goal of acquiring and maintaining situational awareness. Dr Fricker also clearly acknowledges various disciplines and problem areas of research necessary for more successfully addressing these goals.

In addition to adding my emphasis to the value of using simulation for knowledge discovery in public health, I will comment on three areas of considerable importance that affect biosurveillance, (1) the use of multiple sources and types of data and other information, (2) activities surrounding the implementation of electronic medical records (EMRS) and clinical information exchange as it pertains to biosurveillance, and (3) locating the biosurveillance operations focus within the public health structure and organization. Practical necessities and current operational status are large determinants of priorities for biosurveillance and how it may reasonably be implemented.

Simulation

Dr Fricker points out in Section 2.4.3 that there is a dearth of published studies in the biosurveillance literature that use the sound comparative approaches that are common in other fields for evaluating new methods. I share his sense of befuddlement as to why this is the case. Simulation in particular is misunderstood and under-used. I believe that most researchers marginally experienced in this field can appreciate the use of controlled iterative introduction of random error into a simulated data framework that includes observed features from the real data of interest. However, I have also encountered sentiments that 'simulations are not real', and situations where researchers take one set of real data that has been examined for known outbreaks, retrospectively run algorithms on it, find x change-points out of n in a time series, and attribute a detection metric to an algorithm based on a single realization of disease data. These attitudes and activities indicate a lack of penetration in the understanding and appreciation of simulation methods. It seems quite intuitive to consider that before flying a jet, first experience a flight simulator to gain insight on how the jet will respond to various inputs; before performing heart transplant surgery on a live human patient, practice the procedure on live pigs or other animal models; before trusting a new radar device in a live strategic defense event, plan and carry out a few gaming scenarios—making each situation as realistic as possible to ensure that the operator can distinguish threats from 'friendlies' or non-combatants. The field of public health has not yet engendered the production of a variety of public health threat scenarios that range from (1) the very simple and likely to the (2) creatively complex and 'impossible' and embedded them into a Monte Carlo simulation framework. Such a facility would, like the other intuitively reasonable simulated activities mentioned, provide a practice field in which to gain experience about the data types, how people react to methods, the nature of what gets observed in results when changes in information features are introduced at the

Office of Surveillance Epidemiology and Laboratory Services, Public Health Surveillance Program Office, Centers for Disease Control and Prevention, 1600 Clifton Road MS E97, Atlanta, GA 30333, U.S.A.

**Correspondence to: Henry R. Rolka, Office of Surveillance Epidemiology and Laboratory Services, Public Health Surveillance Program Office, Centers for Disease Control and Prevention, 1600 Clifton Road MS E97, Atlanta, GA 30333, U.S.A.*

†*E-mail: HRolka@CDC.Gov, hrr2@cdc.gov*

Published in 2011 by John Wiley & Sons, Ltd.

Statist. Med. ***2011***, 30 416–419

front end, and of course a platform for evaluating detection, display, data management, communication, and reporting methodologies. I vigorously agree with Dr Fricker's assertion about the value to public health for incorporating an approach for evaluating methods that begins with a lucid description of the theory and concepts, then incorporates iterative simulations under various conditions and degrees of randomness to demonstrate practical utility, operational limitations, etc., and concludes with trials within the practical setting in the operational public health milieu.

Multiple types of data and sources of information require new analytic knowledge, skills, and abilities

The distinction between data and information is important for the development of analytic approaches to biosurveillance. In traditional public health surveillance, the term data is often used to refer to record-level detail. Public health still relies strongly on data collected in surveys or record-by-record in surveillance settings. These data are analyzed using time-tried statistical methodologies for exploration and inference in public health. Now there is also a wealth of real-time public health information available for potential surveillance value in the form of unstructured or text data. Data or information in such form includes chief complaints at the record level and news or intelligence-like reports that come from the news media, social network, web-based, and other information aggregation systems. The process of combining information from sources such as these to accomplish situational awareness and new knowledge discovery has been referred to as information fusion. (For further information on this topic and example applications, see [1–3].) In Section 2.6, Dr Fricker refers to the need for additional information from a bioterrorism defense perspective, such as the likely location of an attack. Such additional information would have to come from outside the public health community and the information exchange interface would involve policy and legal as well as technical challenges. Putting the policy and legal issues aside for now, let us further consider *by whom* the technical data analytic challenges will be met.

Approaches and methods for analyzing and understanding record-level data are very different from those used to analyze and understand unstructured information. Additionally, since unstructured data tend to be more anecdotal in nature, and also delivered much more quickly than traditionally sourced surveillance data, the need to combine or fuse data and information of different types adds a layer of complexity to the analytics. The analytic data management skills and tools for record-level, spontaneously generated data from automated systems are very different from those needed to make use of unstructured data. Analysis of both of these types of data requires skill sets that are different from those needed to analyze and make sense out of traditional survey data collected by design. The abundance and availability to public health of both structured and unstructured data sources has grown rapidly over the last decade. Indeed, much faster than the analytic infrastructure has been able to adapt. Currently, the supply of staff with the required knowledge, skills, and abilities in public health for analytic data management and the data analysis itself is far less than the demand. For the interested reader, Davenport and Harris [4] explore reasons for the general institutional lag with respect to analytics. It is indisputable that when there is an information-intensive mission environment and the data analytic component goes unmet, progress suffers.

Although the primary goals and objectives for public health are quite stable, in the emerging area of real-time biosurveillance the information context (types of data, how it is acquired, procedures for processing, the large volumes, demands for rapid cycle analyses, etc.) has evolved very quickly. Hence, analysts who are skilled in biosurveillance have learned primarily from hands-on experience. The empirical characteristics of various data types and means for combining information and reporting findings are distinctive. Even with a sound educational foundation that includes data analysis, strong inductive and deductive reasoning skills, and public health knowledge, it takes staff new to this area a year or so of immersed experience in 'operations' to gain a working understanding of the context. This rapid growth in a field of application that largely resides within public health has created a gap in human resources. For the people who can perform the complex data analysis and even more so, the necessary precursory analytic data management is scarce and urgently needed. The field is in need of modern textbooks and curricula within public health, statistics, biostatistics, and informatics programs to prepare new professionals in the area of analytic data management for biosurveillance and information fusion.

EMRS, health information exchanges (HIES), and biosurveillance

Dr Fricker's manuscript was not intended to address government policies for improving access to health, health record keeping efficiency, and health-care quality. However, some recent government initiatives promoting change in how information is processed from the point of clinical encounter are very likely to have an impact on the methods' focus addressed by this paper. In addition, as the analytical techniques evolve in the health-care delivery industry to utilize EMR data for quality of care assessment and improving patient access to health and information, public health analysts

Published in 2011 by John Wiley & Sons, Ltd.
Statist. Med. ***2011***, 30 416–419

should collaborate with the health-care industry to ensure efficient and effective sharing of new or novel analytical techniques. Blossoming investment in electronic health data systems is certain to lead to an influx of record level and aggregated data that can be used for biosurveillance. The federal government continues to promote EMR and HIE implementations and has invested hundreds of millions of dollars under Recovery Act funding toward '…supporting public health agencies' authorized use of and access to electronic health information; …and such other activities as the Secretary may specify' [5]. Additional sources of readily available spontaneously generated health monitoring data will be available in more standardized forms, with substantial consequences for public health surveillance activities of all kinds, including surveillance of chronic and infectious diseases, immunizations, occupational hazards, pharmaceutical use and risks, environmental exposures, etc. This will affect the data management and analytic requirements that we have described. The demand for data management, analytic methods, and skilled analytic professionals will increase dramatically with the accelerating availability of information-rich health record data. The most intensive stages of this data availability growth have yet to be realized. Intra and inter-institutional connectivity issues (e.g. data standards, harmonization, architectures, etc.) will affect data analytic methods and techniques. Professionals working in the area of biosurveillance data analysis will have an advantage of understanding and producing value if they are aware of how this national data recording system is developing and becoming implemented.

Public health priorities and biosurveillance analytic requirements

There are a multitude of long validated practices and activities that must be carried out within the public health infrastructure at the local, state, and federal levels. These include immunization, infectious disease, chronic disease, injury, birth defect, occupational health, and environmental health programs, just to name a few. Each of them has monitoring or surveillance, as well as programmatic intervention and prevention components. As Dr Fricker has pointed out, biosurveillance and sydromic surveillance became much more prominent following the perceived threat of bioterrorism following 11 September 2001. This represented a new requirement for most public health offices. Along with the new requirement came a great deal of uncertainty. The data, tools, and means for information exchange were not well understood and did not easily fit in with traditional public health practice [6]. Since that time there have been abundant opportunities for trial-and-error learning necessary to account for (1) variations in diagnostic assignments, transcription and coding at the clinical level, (2) data errors introduced in transactional messaging between the clinical environments and public health data systems, (3) parsing of HL7 messages into a relational architecture accessible to public health staff, (4) construction of analytic data marts to be queried by public health staff, (5) production of analytically relevant flat files, and (6) application of statistical models and algorithms for identification of potential signals and/or associations of interest inferred by public health subject matter experts. This complexity is often over-simplified by stating 'sydromic surveillance causes too many false alarms'. The process of incorporating macro-level quality assurance for data across a broad range of information processing environments is a very large task area. It is necessary however to answer the fundamental question: 'How much reality is reflected in biosurveillance data'? The fields of epidemiology, statistics, operations research, etc. have methods for communicating the degree of uncertainty in findings from data *after* this more primary question is addressed. We still have a long way to go in managing the documentation and quality of the biosurveillance data stream. Such work responsibilities reside in the public health domain and will require an influx of human resources skilled, knowledgeable, and motivated to step up to the challenge of more fully characterizing the biosurveillance data space. This should not occur at the expense of time-tried public health practices but in conjunction with their requirements for enhanced and timelier surveillance information. Dr Fricker's work provides reinforcement to an evolving success for using new information sources to enhance public health.

Disclaimer

The assertions and conclusions in this commentary are those of the author and do not necessarily represent the views or position of the Centers for Disease Control and Prevention.

References

1. Rolka H, O'Connor JC, Walker D. Public health information fusion for situation awareness. *Biosurveillance and Biosecurity*, *International Workshop*, *Biosecure 2008 Proceedings*. Lecture Notes in Computer Science. Springer: Germany, 2008; 1–9.
2. Chretien JP, Anyamba A, Small J, Tucker CJ, Britch SC, Linthicum KJ. Environmental biosurveillance for epidemic predictions: experience with rift valley fever. *Biosurveillance and Biosecurity*, *International Workshop*, *Biosecure 2008 Proceedings*. Lecture Notes in Computer Science. Springer: Germany, 2008; 169–174.

Published in 2011 by John Wiley & Sons, Ltd.

3. Zeng DD, Yan P, Li S. Spatial regression-based environmental analysis in infectious disease informatics. *Biosurveillance and Biosecurity*, *International Workshop*, *Biosecure 2008 Proceedings*. Lecture Notes in Computer Science. Springer: Germany, 2008; 175–181.
4. Davenport TH, Harris JG. *Competing on Analytics*. Harvard Business School Press: Boston, MA, 2007.
5. ARRA. The Public Health Service Act as amended by The American Recovery and Reinvestment Act of 2009 See Section 3012 of PL 111-5. Available from: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=111_cong_bills&docid=f:h1enr.txt.pdf[accessed 19 February 2010].
6. Reingold A. If sydromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism*: *Biodefense Strategy*, *Science and Practice* 2003; **1**:1–5.