Reuse Of Clinical Data

C. Safran
Division of Clinical Informatics, Beth Israel Deaconess Medical Center, Harvard Medical School,
Boston, MA, USA

Summary

Objectives: To provide an overview of the benefits of clinical data collected as a by-product of the care process, the potential problems with large aggregations of these data, the policy frameworks that have been formulated, and the major challenges in the coming years.

Methods: This report summarizes some of the major observations from AMIA and IMIA conferences held on this admittedly broad topic from 2006 through 2013. This report also includes many unsupported opinions of the author.

Results: The benefits of aggregating larger and larger sets of routinely collected clinical data are well documented and of great societal benefit. These large data sets will probably never answer all possible clinical questions for methodological reasons. Non-traditional sources of health data that are patient-sources will pose new data science challenges.

Conclusions: If we ever hope to have tools that can rapidly provide evidence for daily practice of medicine we need a science of health data perhaps modeled after the science of astronomy.

Keywords

Electronic health records, database, evidence-based medicine, outcomes research, clinical informatics

Yearb Med Inform 2014:52-4
http://dx.doi.org/10.15265/IY-2014-0013
Published online August 15, 2014

As a byproduct of a patient's care, vast quantities of information are stored in electronic databases. The primary reason for collecting this information is to support the care of the patient during an encounter or subsequent encounters. For the purpose of this review, all other uses of a particular patient's data not for that patient's care will be considered reuse. The reuse of patient data for quality assurance and clinical research is not new, but in the context of "big data", has new importance both for the prospect of refining the "evidence" that we base medical decisions upon as well as the potential for gaining new insights in the era of personalized medicine. This review will highlight some of the benefits of reuse, the potential problems with large clinical databases, the policy frameworks that have been formulated, and the major challenges in the coming years.

The volume and availability of health data has increased primarily for two reasons – the mandated adoption of data exchange standards and the variety of types and sources of data. The stimulus to adopt information technology in healthcare is driven by the belief that it can help control costs as well as improve the safety of care. Demonstrating the improvement in the quality or safety of care is much easier than proving that health information technology saves money. For instance, while automation in the clinical laboratories has improved efficiency, in our hospital we have re-purposed personnel to perform other tasks. In the United States, these drivers are embodied in the HITECH act of 2009 that provides incentives for hospitals and physicians to adopt electronic health records (EHRs) that are interoperable. The trends are similar worldwide.

In three decades (1983 to 2013), the data storage needs of our hospital has increased by about six orders of magnitude – from two gigabytes to approximately two petabytes of data. Although our hospital has merged with another hospital and our EHR now captures all clinical notes in every of the 70 specialty clinics, the increase is largely due to the storage of images. In addition to the character-based data in EHRs and the largely unstructured data of images, we routinely collect vast quantities of data from medical devices in our intensive care units and at patient bedsides. The storage of genomic data still remains an active source of discussion because some have argued we will only need to store variations that are relatively rare in the three billion base-pairs in each patient's DNA, while others argued the full genome should be stored for each patient. Nevertheless, storage of data is not the issue, but rather what should we do with all the data we can now collect?

The dream of those who advocate for the practice of evidence-based medicine [1] is that quality evidence exists to guide clinicians through the clinical conundrums they routinely face – which test to order; how to interpret the test results and what therapy to try? Ideally we would like to find this evidence within the results of a randomized controlled trial (RCT), but we know that RCTs are expensive and cover only a small fraction of clinical situations. Moreover, inclusion and exclusion criteria mean that rarely is the evidence generated by RCTs strictly about patients like my patient [2].

Will humongous (really, really large) databases of routinely collected clinical data from EHRs be an acceptable alternative to find evidence upon which we can base the practice of medicine [3, 4]? Certainly, we need these large data stores to analyze the care of patients with rare conditions. As new pharmaceuticals are brought to market, we need good surveillance to detect adverse effects not detected in clinical trials. Again providing motivation to aggregate routinely collected clinical data beyond the regional domain to national and international contexts.

The benefits of reusing clinical information have been well documented in the clinical literature for decades [5-7]. Cohort analysis has been used to determine risk for readmission to the hospital within 30 days from discharge [8]; predict death and length of stay based upon abnormal laboratory values [9]; describe populations of patients [10, 11]; assist in infection control [12-18]; and discover pharmaco-epidemiological relationships [19, 20].

Clinical information systems, however, are not structured to support *ad hoc* queries, but rather the retrieval of an individual patient's information. Information from EHRs needed to be aggregated into clinical data repositories [21-24], registries [5, 6] and data warehouses to support clinical investigation. New tools were also developed to assist clinicians with the task of query [21, 25-30]. Tools such as i2b2 (Informatics for Integrating Biology and the Bedside) provide an open source infrastructure for aggregating clinical data from multiple sources with graphical tools supporting advanced query and are increasingly used in the international community.

Clinical registries are usually well structured; meaning the sites contributing data use standardized forms and controlled vocabularies. Sites that contribute to a registry have executed data sharing agreements and usually trust the host organizer. While the data within a registry lag behind real-time, researchers have produced a plethora of scientific results. In the case of the registry called ARAMIS (the Arthritis, Rheumatism, and Aging Medical Information System -- previously called the American Rheumatological Association Medical Information System), investigators have published more than 800 articles [31].

In contrast to registries, aggregations of clinical data from EHRs such as the General Practice Research Database (GPRD) [11] in the United Kingdom are less well structured, but are close to real-time. Clinical data are messy and often incomplete or at least irregular. Clinical data are always collected for a purpose and at a cost. Hence, there is inherent bias within routinely collected data. Moreover, the data density across patients is not regular; data are missing. Clinicians choose when to order tests so not every similar patient will have a particular test at a particular time. Clinicians also introduce bias when they select treatments without providing the reason for selection. When data are aggregated from more than one clinical setting, the meaning of data may not be consistent. For instance, if a cardiologist records chest pain on a problem list he might mean something different from a gastroenterologist who has recorded the same problem for a different patient. When data are aggregated, the context of how and why the data were collected may not be transmitted. Lastly, in contrast to registries, clinical data repositories often have more process-related data rather than true outcome data. Because of these inherent biases and limitations, researchers must be cautious when conducting analyses of routinely collected clinical data.

Because real-time aggregation of clinical data presents challenges, some have criticized the analysis of humongous clinical databases as flawed and not reproducible [32]. For instance, many consider a p-value less than 0.05 significant. If a researcher conducts 20 random comparisons, on average, one comparison will reach statistical significance. Imagine instructing a computer program to make millions of comparisons across a database with millions of patient records. Some discovered relationships might be important, but unless other scientists can analyze the same data set to validate the results, the discovered results will not be reproducible. If the database updates itself nightly, even the original researchers might have trouble reproducing their results a year later!

But statistics and data integrity and data consistency are not the greatest challenges facing those wishing to aggregate ever larger sets of patient data across institutions, regions, and nations. As many have said, "culture trumps everything" and many of our citizens do not want to trust large organizations, businesses, or governments with our personal data. Every day in the news we learn of some data breach or worse governmental misuse of our personal data. The American Medical Informatics Association [33, 34] and the International Medical Informatics Association [35, 36] have convened experts from academia, industry and governments to build a framework for trusted stewardship for the reuse of health data. Experts are in broad agreement that sharing clinical data for public health and scientific research should be facilitated by national and transnational policies. Moreover, they agree that the lack of a policy framework risks the safety and health of our citizens [35]. However, the sale of health data remains an unresolved policy issue [33, 36]. Many health organizations are creating sizable clinical data repositories to realize internal value. These same organizations speculate that their data might have value to others who are willing to pay for access. Some health information technology companies even offer a service to broker access to the data of their customers. Technically this could be greatly facilitated by cloud-based EHRs because the data are already centralized, and organizations already have an element of trust with the provider of the cloud-based solutions. While academics and governmental officials often condemn the commercial reuse of routinely collected clinical data, the practice is widespread with governments (at least in the United States) being the largest purchasers of clinical data. Today, fragmented or non-existent national and international policies do not support international aggregation or querying of routinely collected clinical information.

If the technical, scientific and political issues surrounding the reuse of routinely collected clinical data were not daunting enough, the very meaning of clinical data is evolving. Of course, each individual has 3 billion base pairs in their DNA and the availability and use of this genomic information is just beginning. Some have speculated that the human microbiome will prove to be an important predictor of health and wellbeing. But new person-contributed data might also be incorporated into the evolution of the EHR. Sensors in the home and on a person can already generate vast quantities of clinical data. A single continuous glucose monitor can sample interstitial glucose one a minute or over 30,000 times a month! While home-based devices like the continuous glucose monitor have an analogy to the medical devices already in hospital, if every diabetic patient in the United States (25 million) or the world (an estimated 350 million) had such a device, who would look at all this data? Consumers can already report symptoms and side effects of medications on public websites. Soon, this will be part of personal health records and linked if appropriate to EHRs. Thus, new sources of data direct from the healthcare consumer will at least equal if not overwhelm the data we now are aggregating in clinical data repositories.

Clem McDonald once described our challenge of using routinely collected clinical data as analogous to the problems astronomers are facing [3]. The rationale for overcoming the barriers inherent with clinical data seems compelling. In the era of personalized medicine, chronic diseases such as cancer will be re-categorized as multiple rare diseases so that ever-larger datasets will be needed to understand their diversity. Moreover, if we ever hope to develop a "learning health care system" that can rapidly develop evidence for daily practice [37] we need a science of health astronomy and a Hubble telescope for health.

References:

1. Guyatt G, Cairns J, Churchill D, Cook D, Haynes B, Hirsh J, Irvine J et al. Evidence-based medicine. JAMA 1992;2420-25.
2. Safran C. Medicine based upon data. (Editorial) J Gen Intern Med 2013 Dec;28(12):1545-6.
3. McDonald CJ, Hui SL. The analysis of humongous databases: problems and promises. Stat Med 1991 Apr;10(4):511-8.
4. Smith DM. Database research: is happiness a humongous database? Ann Intern Med 1997 Oct 15;127(8 Pt 2):725.
5. Starmer CF, Rosati RA, McNeer JF. Data bank use in management of chronic disease. Comput Biomed Res 1974 Apr;7(2):111-6.
6. Fries JF, McShane D. ARAMIS: a national chronic disease data bank system. In: Proceedings of the Ann Symp on Comp App in Med Care 1979; 798-801.
7. Safran C. Using routinely collected data for clinical research. Stat Med 1991 Apr;10(4):559-64.
8. Phillips RS, Safran C, Cleary PD, Delbanco TL. Emergency readmission for patients discharged from the medical service of a teaching hospital. J Gen Int Med 1987;2:400-5.
9. Herrmann FR, Safran C, Levkoff SE, Minaker KL. Serum albumin on admission as a predictor of death, length of stay, and readmission. Arch Intern Med 1992;152:125-30.
10. Herrmann FR, Safran C. Real-time exploration of routinely collected data: an analysis of admissions for AIDS in a teaching hospital. Medinfo 1992, Proceedings of the Seventh World Conference on Medical Informatics. 1992; 878-82.
11. Hansell, A, Hollowell J, Nichols T, McNiece R, Strachan D. Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). Thorax 1999 May;54(5):413-9.
12. Classen DC, Burke JP. The computer-based patient record: the role of the hospital epidemiologist. Infect Control Hosp Epidemiol 1995;16(12):729–36.
13. Evans RS, Burke JP, Classen DC, Gardner RM, Menlove RL, Goodrich KM, et al. Computerized identification of patients at high risk for hospital-acquired infection. Am J Infect Control 1992; 20(1):4–10.
14. Chizzali-Bonfadin C, Adlassnig KP, Koller W. MONI: an intelligent database and monitoring system for surveillance of nosocomial infections. Medinfo 1995;8:1684.
15. Kahn MG, Steib SA, Fraser VJ, Dunagan WC. An expert system for culture-based infection control surveillance. Proc Annu Symp Comput Appl Med Care 1993:171–5.
16. Pittet D, Safran E, Harbarth S, Borst F, Copin P, Rohner P, et al. Automatic alerts for methicillin-resistant Staphylococcus aureus surveillance and control: role of a hospital information system. Infect Control Hosp Epidemiol 1996;17(8):496–502.
17. Samore M, Lichtenberg D, Saubermann L, Kawachi C, Carmeli Y. A clinical data repository enhances hospital infection control. Proc AMIA Annu Fall Symp 1997;56–60.
18. Geva A, Wright SB, Baldini LM, Smallcomb JA, Safran C, Gray JE. Spread of methicillin-resistant Staphylococcus aureus in a large tertiary NICU: network analysis. Pediatrics 2011;128:e1173-80.
19. Chalasani N, Aljadhey H, Kesterson J, Murray MD, Hall SD. Patients with Elevated Liver Enzymes are Not at Higher Risk for Statin Hepatotoxicity. Gastroenterology 2004;126(5):1287–92.
20. Herzig SJ, Howell MD, Ngo LH, Marcantonio ER. Acid-suppressive medication use and the risk for hospital-acquired pneumonia. JAMA 2009;301(20):2120-8
21. Safran C, Porter D, Lightfoot J, Rury CD, Underhill LH, Bleich HL, et al. ClinQuery: a system for online searching of data in a teaching hospital. Ann Intern Med 1989;111(9):751-6.
22. Walley T, Mantgani A. The UK general practice research database. Lancet 1997 Oct 11;350(9084):1097-9
23. Brown SH, Lincoln MJ, Groen PJ, Kolodner RM. VistA—US Department of Veterans Affairs national-scale HIS. Int J Med Inform 2003;69(2),135-56.
24. Danforth K, Patnode CD, Kapka TJ, Butler MG, Collins B, Compton-Phillips A. Comparative Effectiveness Topics from a Large, Integrated Delivery System. Perm J 2013 Fall;17(4):4-13.
25. Nigrin DJ, Kohane IS. Temporal expressiveness in querying a timestamp-based clinical database. J Am Med Inform Assoc 2000;7(2):152–63.
26. Nigrin DJ, Kohane IS. Data mining by clinicians. Proc AMIA Symp 1998:957–61.
27. Nigrin DJ, Kohane IS. Scaling a data retrieval and mining application to the enterprise-wide level. Proc AMIA Symp 1999:901-5.
28. McDonald CJ, Dexter P, Schadow G, Chueh HC, Abernathy G, Hook J, et al. SPIN query tools for de-identified research on a humongous database. AMIA Annu Symp Proc 2005; 515–9.
29. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 2009 Sep-Oct;16(5):624-30.
30. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010 Mar-Apr;17(2):124-30.
31. http://aramis.stanford.edu (accessed 4/1/2014)
32. Trouble at the Lab: Unreliable Research. Economist October 2013.
33. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14:1-9.
34. Hripcsak G, Bloomrosen M, Brennan P, Chute CG, Cimino J, Detmer DE, et al. Health data use, stewardship, and governance: ongoing gaps and challenges: a report from AMIA's 2012 Health Policy Meeting. J Am Med Inform Assoc 2013 Oct 29. doi: 10.1136/amiajnl-2013-002117.
35. Geissbuhler A, Safran C, Buchan I, Bellazzi R, Labkoff S, Eilenberg K, et al. Trustworthy reuse of health data: a transnational perspective. Int J Med Inform 2013;82:1-9.
36. Bellazzi R, Buchan IE, Labkoff SE, Geissbuhler A, Safran C. The IMIA initiatives on trustworthy reuse of health data: A report. Stud Health Technol Inform 2013; 192:1231.
37. Etheredge LM. A rapid-learning health system. Health Aff (Millwood) 2007 Mar-Apr;26(2):w107-18.

Correspondence to:
Charles Safran, MD
Division of Clinical Informatics
Beth Israel Deaconess Medical Center
Harvard Medical School
Boston, MA, USA
E-mail: Charles_Safran@Harvard.Edu