Update on Data Reuse in Health Care

C. Safran
Harvard Medical School, Boston, United States

Summary

Objective: Reuse of clinical data has broad use in clinical, research, governmental, and business settings. This summary provides an update on the benefits, barriers to use with large clinical databases, policy frameworks that have been formulated, and challenges.

Methods: This report highlights some recent publications on the diverse uses of clinical data and some policy initiatives to promote reuse. It also contains the opinions of the author.

Results: Although many examples of the benefits of data reuse have been documented, this summary also reviews why the quality of clinical data needs to be the focus of future informatics work.

Conclusion: The promise of reusing data outweighs potential risks, but concerns about privacy and the need to modernize our legal framework will be necessary to realize the full benefits of real-world evidence.

Keywords

EHR; electronic health records; big data; health policy; learning health systems; real world evidence

Yearb Med Inform 2017:24-7
http://dx.doi.org/10.15265/IY-2017-013
Published online September 11, 2017

1 Introduction

The use of data for reasons other than originally intended is broadly termed *secondary use* or, more appropriately, *reuse.* The reuse of clinical data is not new and has been the basis for advancement in the science of medicine and the processes of health care [1, 2]. However, the widespread adoption of technologies such as electronic health records (EHRs), new sources of health information [3], and the training of clinical informaticians [4], has accelerated the interest in and importance of this topic. This survey will review recent areas of clinical data reuse as well as will highlight areas of concern and future work.

2 Actors of Reuse

Research and Patient Care

Reuse of clinical data is often considered in the context of research, but this data has broad use in clinical, governmental and business settings, including but not limited to those in Table 1. The broad vision, termed "the learning health system", implies that we can use the data collected as a by-product of clinical care (e.g., big data for health care) to improve the performance of our health care system and provide individual patients with the best possible information about their diagnostic and treatment choices [3, 5, 6]. Repositories of clinical data have been used for direct patient care to support the finding of similar patients. Such systems can support queries such as "Have we ever seen a patient like this before, and if so what was the diagnosis and what treatments were given?" [7]. Ultimately, treatments should be linked to clinical outcomes so that the best treatment can be selected. Increasingly, traditional sources of health-related data, such as claims databases, can be linked with clinical data from EHRs to better understand the safety of medications we prescribe to patients [8]. In fact, the wealth of information in all sources of data can be used to improve efficiency in clinical care [9] as well as compare the effectiveness of therapeutic options [10].

Reuse of data for research may have the most far-reaching impact on the health of our citizens by not only speeding the design and execution of clinical trials [11] but also assisting in the discovery of new knowledge [12-14].

Table 1 Some Common Uses of Data

Clinical Care
Direct Patient Care
Quality and Safety
Improving Efficiency
Comparing Effectiveness
Population Health
Governmental
Biosurveillance
Immunization Tracking
Developing Public Policy
Business
Fraud Detection
Calculation of Insurance Premiums and Risk
Marketing & Sales
Drug Development
Post-Marketing Surveillance
Research
Design of Trials
Recruitment
Prediction
Discovery

Government

In the United States, the government is the largest provider of health care, and therefore it is the largest consumer of health data for secondary reasons. Both at federal and local levels, public health entities have the authority to collect and analyze health data [15]. Claims databases, which can include all federally funded health care claims, have been used to develop heath care policy [16].

Business

Health care is an industry, and data from and about its multi-trillion-dollar economy has monetary value. The United States government has incentivized physicians and hospitals to adopt EHRs not only to improve care but also to decrease billing fraud [17]. Insurance companies can use their claims databases to perform actuarial calculations that estimate risk for patient populations so they can set insurance premiums. Companies such as QuintilesIMS generated profits of almost $400 million from the acquisition, analysis, and sale of information derived from clinical data. Not surprisingly, pharmaceutical companies are some of QuintilesIMS's best customers.

Historically, pharmaceutical companies needed post-marketing data to compensate their sales force. In marketing a pharmaceutical product, a salesperson visits a physician's office or hospital in a geographic location in the hope of influencing future sales. One way a pharmaceutical company can understand the effectiveness of a salesperson is to monitor regional sales from databases provided by companies like QuintilesIMS. Marketing campaigns, including those annoying "ask your doctor" TV ads, are measured by analyzing these same databases. These databases can also be linked to claims data, providing additional insight as to which diagnoses are associated with which pharmaceutical products. Moreover, laboratory data linked to these same datasets can be used for post-marketing surveillance of potential adverse events. Once genomic data will be integrated into EHRs, one can assume that pharmaceutical companies will want to purchase this data as well.

3 Constraints for Reuse

Collecting Data as a By-product of Care

The quality and accuracy of collected data is critical for data reuse. Data collection has a cost, and since the purpose at the time of collection is supporting the clinical workflow, complying with federal regulations, and increasing reimbursements, the quality of the data for reuse is not a priority. Clinical informaticians must engage EHR vendors and clinicians in their institutions to ensure that the data collected is stored in a way which will support downstream reuse.

During the implementation process, a health system can choose how much unstructured dictation to allow and which terminologies (SNOMED, ICD, UMLS, etc.) to use for structured data. The tension in clinical system implementation involves the balance between requiring structured information versus capturing clinical narrative information in unstructured form. Most data in EHRs are unstructured and, to date, natural language processing has not reliably been used to structure this wealth of data.

To capture a single clinical fact in a fashion that is reproducible takes between 5 and 7 seconds [18]. In the clinical realm, the time allotted to documentation is already a burden for our providers, and the regulatory environment, including "meaningful use," is taxing clinicians' good will to the limits [19]. Consequently, most EHR implementations structure only those data required by regulation or for reimbursement.

Loss of Information

Getting the data "right" is a challenge in every clinical setting since data must be acquired, sometimes transformed (e.g., coded), and eventually stored. In a clinical system, there are hundreds of thousands of variables. Even facts like a patient's serum sodium level can be associated with many elements or *metadata*, such as when the test was ordered; for whom the test was ordered; why it was ordered; when and by whom the blood sample was drawn; when the sample of blood arrived at the laboratory; how, when, and by whom it was processed; what the normal range is for that sample; who looked at the result and when; etc. Some facts of care, such as what medications a person is taking, are even more complicated than a single laboratory result. In a hospital setting, there is both a pharmacy record of what is ordered and dispensed and a nursing medication administration record. Pills are relatively easy to count, whereas intravenous admixtures are more difficult to document. Once a patient leaves the hospital, the accuracy of which prescribed medications are actually taken becomes less clear. Finally, for each clinical encounter, major diagnoses and procedures are recorded, usually as ICD10 and CPT codes. The inaccuracy of coded diagnoses has been well documented [20].

Data warehouses and repositories used for research store only a tiny fraction of the data collected by clinical systems. Data stored within clinical systems and EHRs are optimized for the display of a single record at a time. Aggregation of this data to answer questions such as "Did the diabetic control of a population of patients improve?" is time consuming and might interfere with clinical operations. Eventually a subset of data from each clinical encounter could be transferred into a disease-specific registry (e.g., cancer registry) or into a clinical data warehouse for future reuse [7, 20, 21]. This selection of only specific data elements for inclusion in a registry or warehouse results in a loss of information and context. For instance, when a physician records on a problem list that a patient has chest pain, a single code is sent to the data warehouse. The name of the physician or the physician's specialty training is typically not associated with this code. The result is that this diagnosis code is treated by researchers as having a single meaning across all instances of it, regardless of who entered it, even though the meaning of chest pain may be different when specified by a gastroenterologist or by a cardiologist.

Need for Aggregation

The volume of data from a single clinical setting is seldom sufficiently robust to provide answers to questions outside that specific clinical setting. Particularly for rare outcomes or

rare conditions, the need for aggregation from multiple settings is imperative. Some initiatives such as PCORI in the United States [22], and EuroRec in Europe [23], are beginning to address this need. However, barriers exist at regional, national, and transnational levels to the reuse of clinical data. These barriers include ownership or rights to use data, the appropriateness of methods to analyze the data, the propriety of the question being analyzed, the legal context for the analysis, and even the underlying language of the data. The American Medical Informatics Association and the International Medical Informatics Association have held a series of policy-related meetings to provide a framework for discussion of these complex issues [24-27].

4 Newer Sources of Health Data

Person Generated Health Data

More health-related data exists outside the health system than within the system. Internet-connected computer games like Nintendo's Wii Fit routinely collect activity data as do most smart phones. Apps on smart phones allow the user to perform tasks such as documenting daily food intake, tracking symptoms, recording daily vital signs, and even recording when medications are taken. Other apps such as LibreLink interface with external sensors to track conditions such as diabetes. Diabetic patients can continuously monitor interstitial glucose on a minute-to-minute basis, leading to thousands of measurements a month.

Genomic and other -omics Data

Genetic sequencing and the related ability to elucidate the human biome are producing important data at a rate and cost-point unimagined only a few years ago. Most EHRs can only include these types of data in an unstructured fashion and hence they are hard to use in population studies. Future aggregation of -omics data will require not only a standard representation but also new legal and ethical guidelines for reuse.

Sensors

Perhaps the greatest contributor to a future healthcare data tsunami will be sensors in the home and on a person. A GPS-enabled watch can track location and activity. These same sensors can continuously monitor heart rate, respirations, and even sleep patterns. Motion detectors in the home can establish a daily activity pattern which can be used to monitor deviations from normal.

5 Ethical Considerations

Privacy dominates the discussion of the reuse of health data. Health data is valuable and increasingly the subject of cyber-attacks. The Health Information Portability Accountability Act (HIPPA) of 1996 is the law in the United States that governs the sharing of patient-identifiable data. For researchers to use clinical data, their institutional review board (IRB) must oversee their use of the data. Although a patient might have the legal right to see his or her medical record or possess a copy of the record, hospitals and office-based physician practices consider the medical record and the data contained within the record to be their property. Patients are not informed when their data are reused, and they do not receive any compensation when their data are monetized.

HIPAA only applies to covered entities (such as hospitals and physicians), but surprisingly many organizations (such as life insurance companies, employers, companies like Facebook and most schools) are not covered. For instance, when applying for life insurance, a person is required to allow the company to look at his or her health records. A summary of health-related information (outside of HIPAA constraint) is maintained by a company owned by these providers [28]. If the person objects to this anonymous aggregation of data, he or she simply is denied coverage.

With the permission of the end user, mobile apps and social media sites such as Facebook can track activity and health data. The lengthy user agreements are rarely read, and if the user does not agree, then he or she cannot use the app or service. The companies behind these services often do not tell users how their data are used.

While the legal framework in Europe and elsewhere in the world is not the same as in the United States, ethical challenges persist. Health data are valuable, and, while governments try to protect their citizens, the technology is moving faster than the regulatory environment. This is evident for genomic data – since genes are inherited, a parent's genomic data creates a privacy risk for their children. Created in 1996, before the Human Genome project was launched in 1998, HIPAA does not address any aspect of inheritance.

6 Conclusion

The promise of reusing data collected as a by-product of care processes and combining this data with the tsunami of health-care-related data coming from outside our health care institutions will transform the practice of medicine and the delivery of health care. Experts agree that the benefits to society of data reuse outweigh potential risks [26]. However, issues of privacy and re-identification of personal information can shift any public discussion. The harmonization of public policy and the modernization of our legal framework will be necessary to realize the full benefits of real-world evidence. Finally, monetization of health data remains a difficult topic for discussion where academics and privacy advocates consider data should be freely accessible for trusted reuse, and business interests have already created an industry within the bounds of law.

Conflicts: Dr. Safran is a director and shareholder in Intelligent Medical Objects and has been a consultant to Cerner Corporation and Allscripts Corporation.

Funding: Dr. Safran is supported from a grant from the Agency of Health Care Quality and Research (AHRQ), U.S. Department of Health and Human Services (HHS) R01-HS021495. The opinions expressed in this document do not reflect the official position of AHRQ or the U.S. Department of HHS.

References

1. Safran C. Reuse of Clinical Data. Yearb Med Inform 2014:52-4.
2. Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. Yearb Med Inform 2016:211-8.
3. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA 2014 Jun 25;311(24):2479-80.
4. Gundlapalli AV, Greaves WW, Kesler D, Murray P, Safran C, Lehmann CU. Clinical Informatics Board Specialty Certification for Physicians: A Global View. Stud Health Technol Inform 2015;216:501-5.
5. Learning Health Community. www.learninghealth.org (accessed December 8, 2016).
6. Mandl KD, Kohane IS, McFadden D, Weber GM, Natter M, Mandel J, et al. Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): architecture. J Am Med Inform Assoc 2014 Jul-Aug;21(4):615-20.
7. Safran C, Porter D, Lightfoot J, Rury CD, Underhill LH, Bleich HL, Slack WV. ClinQuery: A System for Online Searching of Data in a Teaching Hospital. Ann Intern Med 1989;111:751-6.
8. Lin KJ, Schneeweiss S. Considerations for the Analysis of Longitudinal Electronic Health Records Linked to Claims Data to Study the Effectiveness and Safety of Drugs. Clin Pharmacol Ther 2016;100:147-59.
9. Stadler JG, Donlon K, Siewert JD, Franken T, Lewis NE. Improving the Efficiency and Ease of Healthcare Analysis through Use of Data Visualization Dashboards. Big Data 2016;4(2):129-35.
10. Ananthakrishnan AN, Cagan A, Cai T, Gainer VS, Shaw SY, Savova G, et al. Comparative Effectiveness of Infliximab and Adalimumab in Crohn's Disease and Ulcerative Colitis. Inflamm Bowel Dis 2016 Apr;22(4):880-5.
11. Kreuzthaler M, Schulz S, Berghold A. Secondary Use of Electronic Health Records for Building Cohort Studies through Top-Down Information Extraction. J Biomed Inform 2015;53:188-95.
12. Chen B, Butte AJ. Leveraging Big Data to Transform Target Selection and Drug Discovery. Clin Pharmacol Ther 2016;99:285-97.
13. Ananthakrishnan AN, Cagan A, Cai T, Gainer VS, Shaw SY, Churchill S, et al. Statin Use Is Associated with Reduced Risk of Colorectal Cancer in Patients with Inflammatory Bowel Diseases. Clin Gastroenterol Hepatol 2016 Jul;14(7):973-9.
14. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW et al. Real-World Evidence – What Is It and What Can It Tell Us? N Engl J Med 2016;375:2293-7.
15. Thomas M. Accomplishments and Opportunities in Biosurveillance. J Public Health Manag Pract 2016;22;S81-2.
16. Freedman JD, Green L, Landon BE. All-Payer Claims Databases — Uses and Expanded Prospects after Gobeille. N Engl J Med 2016;375:2215-7.
17. Kruse CS, Goswamy R, Raval YJ, Marawi S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. JMIR Med Inform 2016;4:e38.
18. Slack WV, Leviton A, Bennett SE, Fleischmann KH, Lawrence RS. Relationship between Age, Education, and Time to Respond to Questions in a Computer-Based Medical Interview. Comput Biomed Res 1988;21:78-84.
19. Wachter RM. The Digital Doctor: Hope, Hype, and Harm at the Dawn of Medicine's Computer Age. McGraw Hill Professional; 2015.
20. Danciu I, Cowan JD, Basford M, Wang X, Wang X, Saip A, Osgood S , et al. Secondary Use of Clinical Data: The Vanderbilt Approach. J Biomed Inform 2014 Dec;52:28-35.
21. 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;17:124-30.
22. Patient-Centered Outcomes Research Institute. PCORnet: The National Patient-Centered Clinical Research Network. Washington, DC: Patient-Centered Outcomes Research Institute; 2015. www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network (accessed Dec 8, 2016).
23. EuroRec Institute. www.eurorec.org (accessed Dec 8, 2016).
24. 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.
25. Bloomrosen M, Detmer D. Advancing the Framework: Use of Health Data — A Report of a Working Conference of the American Medical Informatics Association. J Am Med Inform Assoc 2008;15:715-22.
26. 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 Jan;82(1):1-9.
27. 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 2014 Mar-Apr;21(2):204-11.
28. The Facts about MIB. https://www.mib.com/facts_about_mib.html (accessed March 15, 2017).

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