Big Bad Data: Law, Public Health, and Biomedical Databases

Sharona Hoffman and Andy Podgurski

The accelerating adoption of electronic health record (EHR) systems will have profound impacts on clinical care. It will also have far-reaching implications for public health research and surveillance, which in turn could lead to changes in public policy, statutes, and regulations. The public health benefits of EHR use can be significant. However, researchers and analysts who rely on EHR data must proceed with caution and understand the potential limitations of EHRs.

Much has been written about the risk of EHR privacy breaches.[1] This paper focuses on a different set of concerns, those relating to data quality. Unlike clinical trial data, EHR data is not recorded primarily to meet the needs of researchers. Because of clinicians' workloads, poor user-interface design, and other factors, EHR data is surprisingly likely to be erroneous, miscoded, fragmented, and incomplete. Although EHRs eliminate the problem of cryptic handwriting, other kinds of errors are more common with EHRs than with paper records. Moreover, automated processing of EHR data can eliminate some opportunities for checks by humans. In addition, if causation is at issue, analysts must grapple with the complexities of making causal inferences from observational data. Public health findings can be tainted by the problems of selection bias, confounding bias, and measurement bias. These and other obstacles can easily lead to invalid conclusions and unsound public health policies.

The paper will highlight the public health uses of EHRs. It will also probe the shortcomings of EHR information and the challenges of collecting and analyzing it. Although some of the problems we discuss apply to all research, including that based on paper records, they will become all the more troubling and important in an era of electronic "big data," in which massive amounts of data are processed automatically, without human checks. Finally, we outline several regulatory and other interventions to address data analysis difficulties.

Public Health Benefits of EHRs

The advent of EHRs brings with it a wealth of opportunities for enhanced public health initiatives. EHR systems can report timely data that could facilitate surveillance of infectious diseases, disease outbreaks, and chronic illnesses. Software can extract data from records, analyze them, and electronically submit them to public health authorities, which will likely soon receive unprecedented amounts of information.[2] In fact, the "Meaningful Use" regulations with which providers must comply in order to be granted federal incentive payments for EHR adoption already require that providers be able to submit three types of data to public health authorities: lab results, syndromic surveillance, and immunizations.[3]

EHRs will also greatly facilitate public health research. Large EHR databases can enable researchers to conduct comprehensive observational studies that include millions of records from patients with diverse demographics who are treated in real clinical settings over many years. Researchers could use

Sharona Hoffman, J.D., LL.M., *is the Edgar A. Hahn Professor of Law and Professor of Bioethics and the Co-Director of the Law-Medicine Center at Case Western Reserve University School of Law. She received her B.A. from Wellesley College; J.D. from Harvard Law School; and LL.M. in Health Law from the University of Houston.* **Andy Podgurski, Ph.D.,** *is a Professor of Electrical Engineering and Computer Science at Case Western Reserve University. He received his B.S., M.S., and Ph.D. degrees from the University of Massachusetts.*

these rich collections of data to study disease progress, health disparities, clinical outcomes, treatment effectiveness, and the efficacy of public health interventions, and their findings may influence many public health decisions. To this end, the Patient Protection and Affordable Care Act of 2010 embraces the concept of "comparative effectiveness research" and supports the use of observational studies to evaluate and compare health outcomes.[4]

EHRs may be particularly valuable during public health emergencies. EHR systems may enable responders to obtain critical medical information about disaster victims in the absence of access to their physicians' offices and in the face of local computer failures.[5] Basic EHR systems can also be deployed at disaster scenes or in field hospitals to facilitate data sharing, decision-making, and efficient administrative operations.[6]

Equally beneficial are EHR alert and decision support mechanisms that could serve as a continuous communication channel between clinicians and public health authorities. Public health officials could provide electronic updates and recommendations to clinicians both during emergencies and in ordinary times.[7]

EHR Shortcomings

The proliferation of available data is generating much excitement in the public health community. However, this enthusiasm must be tempered by recognition of the potential limitations of EHR data.

EHRs often contain data entry errors, in part because they can increase physicians' documentation burden. Busy clinicians sometimes type quickly and invert numbers, place information in the wrong patient's record, click on incorrect menu items, or copy and paste narrative from prior visits without carefully editing and updating it.[8]

Much of the information in EHRs is coded using not only the International Classification of Diseases (ICD-9) but also customized lists incorporated into EHR products, and coding can introduce further errors. Codes may be confusing, misleading or too general to indicate the specifics of patients' conditions.[9] Furthermore, EHRs may not accommodate detailed and nuanced natural language notes about patients' medical histories and diagnostic findings.[10]

Commentators have noted that providers collect data for clinical and billing purposes rather than for public health reasons. Thus, EHR content is not always well-suited for public health uses. Furthermore, clinicians may have incentives to "upcode" in order to maximize charges, and this practice can systematically compromise the accuracy of many records.[11] The menus and lists built into EHR systems may facilitate upcoding by suggesting items for which physicians should bill and making it easy to click boxes for charge purposes.

In some instances, EHRs are incomplete, lacking essential information such as treatment outcomes. Patients who receive medication from their doctors often do not report whether the therapy was effective. The absence of return visits may mean that the patients were cured, but it could also indicate that they failed to improve or deteriorated and decided to visit different doctors or specialists.[12]

In addition, patient records are often fragmented. A patient may see multiple doctors in different facilities, and if these practices do not have interoperable EHR systems, pieces of the individual's record will be scattered in different locations. Such fragmentation can hinder surveillance and research efforts because the patient's medical history cannot easily be put together into a comprehensive whole.[13]

EHR vendors are making slow progress towards achieving interoperability, the ability of two or more systems to exchange information and to operate in a coordinated fashion. In 2010 only 19% of hospitals exchanged patient data with providers outside their own system.[14] Vendors may have little incentive to produce interoperable systems because interoperability might make it harder to market products as distinctive and easier for clinicians to switch to different EHR products if they are dissatisfied with the ones they purchased.

The lack of interoperability in EHR systems can also impede data harmonization. Different systems may use different terminology to mean the same thing or the same terminology to mean different things. For example, the abbreviation "MS" can mean "mitral stenosis," "multiple sclerosis," morphine sulfate," or "magnesium sulfate."[15] If the term's meaning is not clear from the context, then analysts may not be able to interpret it correctly.

Analytical Challenges and Causal Inference

Even if the EHR data themselves are flawless, analysts must grapple with a variety of analytical challenges. These may be particularly pronounced in the case of studies seeking to answer causal questions, such as whether certain public health interventions have had a positive impact.[16] EHR data is generally observational, not experimental, and hence treatments and exposures are not assigned randomly. This makes it much more difficult to ensure that causal inferences are not distorted by systematic biases. Analysts and users of research data must be familiar with the risks of selection bias, confounding bias, and measurement bias.

Selection bias can occur when analysts unknowingly employ a study group that is not representative of the population of interest. The group studied might have atypical clinical, demographic, or genetic attributes, and therefore, it would be inappropriate to generalize study conclusions to the population at large.[17]

Confounding bias is a systematic error that occurs because there exists a common cause of the treatment/exposure variable and the outcome variable.[18] For example, socioeconomic factors may be confounders because low income may cause individuals to choose sub-optimal, inexpensive treatments and may also separately lead to deteriorated health because of stress or poor nutrition. A failure to account for socioeconomic status may thus skew study results.

> Secondary use of EHR data in order to promote public health can be facilitated through a variety of approaches. Interoperability, improved infrastructure, and appropriate data analysis techniques are all important contributing factors.

Measurement biases are generated by errors in measurement and data collection resulting from faulty equipment or software or from human error. In addition, patients may provide clinicians with incorrect information regarding their medical histories, symptoms, or treatment compliance because they are confused, have impaired memories, or are embarrassed to tell the truth.[19] Naturally, measurement bias can taint analytical results. Systematic errors, which can arise because of EHR-exacerbated problems such as upcoding, are especially challenging.

Adequate Infrastructure

EHR information that is submitted pursuant to the meaningful use regulations may soon inundate public health agencies. It is entirely unclear that these agencies have the infrastructure to receive, store, process, analyze, and make sense out of the data that is submitted. According to one source, only 15% of states with general communicable disease surveillance systems were able to receive EHR data, and other commentators have noted inadequacies in computing resources and shortages of qualified public health analysts.[20] Having large volumes of electronic information available will not promote public health if the government does not have the capacity to process it and apply the findings it yields.

Recommendations

Secondary use of EHR data in order to promote public health can be facilitated through a variety of approaches. Interoperability, improved infrastructure, and appropriate data analysis techniques are all important contributing factors.

Interoperability

Establishing interoperability and data harmonization is of critical importance to the success of the EHR initiative in general and to its positive impact on public health in particular. Semantic interoperability is defined as the ability to interpret and effectively use exchanged information, achieved through "shared data types, shared terminologies, and shared codings."[21]

As discussed above, vendors may not be eager to support interoperability on their own, and the absence of this capacity remains a major concern in the health care community.[22] Consequently, vendors should be incentivized or compelled to produce interoperable EHR systems. One option is to include semantic interoperability requirements in forthcoming Stage 3 Meaningful Use regulations.

Data Collection and Storage

Interoperability alone, however, will not be sufficient to leverage EHRs for public health uses. Health information technology experts will need to develop software that can scan clinicians' EHRs, extract relevant data, analyze it, and communicate findings in the appropriate format to public health agencies. Such efforts are already underway, as illustrated by the example of the Electronic Medical Record Support for Public Health surveillance platform, described in a recently published paper.[23] Furthermore, to the extent that EHRs do not organically contain all of the information that public health authorities will need, vendors should add forms and fields to their systems that will ask clinicians to capture and enter the necessary information.

In addition, the federal government should provide public health departments with funding to enhance their infrastructure in order to receive and process EHR data. Admittedly, however, the current financial climate may make this recommendation more aspirational than realistic.

Data Analysis

Because the quality of EHR data is variable, analysts should take steps to estimate error rates and characterize uncertainty about data accuracy. The data originators, i.e., clinicians, are in the best position to assess data quality because they can audit a sample of EHRs and verify whether information is accurate by interviewing or examining patients. Public health authorities will receive information from numerous providers and will not have access to patients. Therefore, their ability to assess data quality will be limited. Nevertheless, they may be able to compare data sets from different sources, identify values that appear anomalous, and ask the data originators to investigate their accuracy.[24]

Public health personnel should keep abreast of developments in the rapidly evolving field of causal inference research, such as contemporary theoretical work concerning identification of causal parameters based on properties of causal diagrams. Causal diagrams are used in the disciplines of biostatistics, epidemiology, and computer science. These diagrams consist of points representing different variables, such as treatment, outcome, and other factors (clinical, demographic, genetic, etc.) that should be considered, and the points are connected by arrows, representing causal relationships. Figure 1 is a very simple causal diagram that depicts the relationships among three variables: treatment, outcome, and a confounder. The confounder is a variable, such as severity of illness, that might independently affect treatment choice and outcome and thus should be controlled for. In creating causal diagrams, analysts are compelled to articulate their assumptions about causal relationships between variables and to try to identify all elements that might affect the outcome of interest. The diagrams constitute maps of cause and effect relationships that enable researchers to construct sound statistical models, avoid confounding, and correctly interpret data. Recent causal inference scholarship elucidates how causal parameters can be identified and estimated with the help of computer analysis of complex causal diagrams.[25] An understanding of such developments in causal inference methodology could enhance public health authorities' ability to evaluate research outcomes for purposes of changing or implementing public health policies.

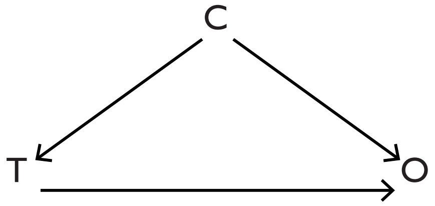

Figure 1

Causal Diagram Showing Causal Arrows between Treatment Variable T, Outcome Variable O, and Confounder C

Conclusion

The transition from paper medical records to EHR systems could have significant benefits for public health. However, public health researchers and surveillance authorities must recognize the potential shortcomings of EHR data and understand how difficult it is to use them to infer causal effects correctly. The public health community should embrace initiatives to leverage EHRs to promote public health, but should approach these with a realistic understanding of the obstacles and challenges they pose.

References

1. See, e.g., L. M. Lee and L. O. Gostin, "Ethical Collection, Storage, and Use of Public Health Data: A Proposal for a National Privacy Protection," *JAMA* 302, no. 1 (2009): 82-84; J. O'Connor and G. Matthews, "Informational Privacy, Public Health, and State Laws," *American Journal of Public Health* 101, no. 10 (2011): 1845-1850; A. Wilson, Note, "Missing the Mark: The Public Health Exception to the HIPAA Privacy Rule and Its Impact on Surveillance Activity," *Houston Journal of Health Law & Policy* 9, no. 1 (2008): 131-156.
2. J. Chretien, N. E. Tomich, J. C. Gaydos, and P. W. Kelley, "Real-Time Public Health Surveillance for Emergency Preparedness," *American Journal of Public Health* 99, no. 8 (2009): 1360-1363; P. F. Smith, J. L. Hadler, M. Stanbury, R. T. Rolfs, R. S. Hopkins, and the CSTE Surveillance Strategy Group, "'Blueprint Version 2.0': Updating Public Health Surveillance for the 21st Century," *Journal of Public Health Management Practice* (2012) (Epub ahead of print), at 5; M. Klompas, M. Murphy, J. Lankiewicz, J. McVetta, R. Lazarus, E. Eggleston, P. Daly, P. Oppendisano, B. Beagan, C. Kirby, and R. Platt, "Harnessing Electronic Health Records for Public Health Surveillance," *Online Journal of Public Health Informatics* 3, no. 3 (2011): 1-7.
3. 45 C.F.R. §170.205(c)-(e) (2011); Public Health Information Network, *Meaningful Use Fact Sheet: Syndromic Surveillance*, at <http://www.cdc.gov/phin/library/PHIN_Fact_Sheets/FS_MU_SS.pdf> (last visited January 9, 2013).
4. S. Cousens, J. Hargreaves, C. Bonell, B. Armstrong, J. Thomas, B. R. Kirkwood, and R. Hayes, "Alternatives to Randomisation in the Evaluation of Public-Health Interventions: Statistical Analysis and Causal Inference," *Journal of Epidemiology and Community Health* 65, no. 7 (2011): 576-581; T. W. Guilbert, B. Arndt, J. Temte, A. Adams, W. Buckingham, A. Tandias, C. Tomasallo, H. A. Anderson, and L. P. Hanrahan, "The Theory and Application of UW e-Health-Phinex, A Clinical Electronic Health Record-Public Health Information Exchange," *Wisconsin Medical Journal* 111, no. 3 (2012): 124-133, at 124-125; S. Hoffman and A. Podgurski, "Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research," *SMU Law Review* 65, no. 1 (2012): 85-144, at 97-102. The latter article discusses the benefits of observational research and its limitations compared to randomized clinical studies. See also 42 U.S.C. §1320e (2010).
5. S. H. Brown, L. F. Fischetti, G. Graham, J. Bates, A. E. Lancaster, D. McDaniel, J. Gillon, M. Darbe, and R. M. Kolodner, "Use of Electronic Health Records in Disaster Response: The Experience of Department of Veterans Affairs After Hurricane Katrina," *American Journal of Public Health* 97, Supp. no. 1 (2007): S136-S141.
6. G. DeMers, C. Kah, C. Buono, T. Chan, P. Blair, W. Griswold, P. Johansson, O. Chipara, and A. Nilsson, "Secure Scalable Disaster Electronic Medical Record and Tracking System," *2011 IEEE International Conference on Technologies for Homeland Security (HST)* (2011): 402-406; G. Levy, N. Blumberg, Y. Kreiss, N. Ash, and O. Merin, "Application of Information

Technology within a Field Hospital Deployment Following the January 2010 Haiti Earthquake Disaster," *Journal of the American Medical Informatics Association* 17, no. 6 (2010): 626-630.

7. N. Garrett, N. Mishra, B. Nichols, C. Staes, C. Akin, and C. Safran, "Characterization of Public Health Alerts and Their Suitability for Alerting in Electronic Health Record Systems," *Journal of Public Health Management Practice* 17, no. 1 (2011): 77-83; J. Lurio, F. P. Morrison, M. Pichardo, R. Berg, M. D. Buck, W. Wu, K. Kitson, F. Mostashari, and N. Calman, "Using Electronic Health Record Alerts to Provide Public Health Situational Awareness to Clinicians," *Journal of the American Medical Informatics Association* 17, no. 2 (2010): 217-219.
8. T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary Use of EHR: Data Quality Issues and Informatics Opportunities," *AMIA Summits on Translational Science Proceedings 2010* (2010): 1-5.
9. S. T. Liaw, J. Taggart, S. Dennis, and A. Yeo, "Data Quality and Fitness for Purpose of Routinely Collected Data – A General Practice Case Study from an Electronic Practice-Based Research Network (ePBRN)," *AMIA Annual Symposium Proceedings* 2011 (2011): 785-794, at 789; see Botsis *(id.)*.
10. R. Kukafka, J. S. Ancker, C. Chan, J. Chelico, S. Khan, S. Mortoti, K. Natarajan, K. Presley, and K. Stephens, "Redesigning Electronic Health Record Systems to Support Public Health," *Journal of Biomedical Informatics* 40, no. 4 (2007): 398-409, at 405.
11. See Smith, *supra* note 2, at 5; C. S. Brunt, "CPT Fee Differentials and Visit Upcoding Under Medicare Part B," *Health Economics* 20, no. 7 (2011): 831-841.
12. C. D. Newgard, D. Zive, J. Jui, C. Weathers, and M. Daya, "Electronic Versus Manual Data Processing: Evaluating the Use of Electronic Health Records in Out-of-Hospital Clinical Research," *Academic Emergency Medicine* 19, no. 2 (2012): 217-227, at 225.
13. C. C. Diamond, F. Mostashari, and C. Shirky, "Collecting and Sharing Data for Population Health: A New Paradigm," *Health Affairs* 28, no. 2 (2009): 454-66, at 456-457; J. W. Beasley, T. B. Wetterneck, J. Temte, J. A. Lapin, P. Smith, J. Rivera-Rodriguez, and B. T. Karsh, "Information Chaos in Primary Care: Implications for Physician Performance and Patient Safety," *Journal of the American Board of Family Medicine* 24, no. 6 (2011): 745-751, at 747.
14. C. Terhune, "U.S. Pushes Healthcare Providers to Share Records Electronically," *Los Angeles Times*, March 10, 2012, *available at* <http://articles.latimes.com/2012/mar/10/business/la-fi-health-tech-20120310> (last visited January 9, 2013). E. H. Shortliffe and J. J. Cimino, eds., *Biomedical Informatics: Computer Applications in Health Care and Biomedicine* (New York: Springer, 2006): at 952 (defining interoperability).
15. C. G. Chute, "Medical Concept Representation," in H. Chen S. S. Fuller, C. Friedman, and W. Hersh, eds., *Medical Informatics: Knowledge Management and Data Mining in Biomedicine* (New York: Springer-Verlag 2005): at 170; M. R. Gold, C. G. McLaughlin, K. J. Devers, R. A. Berenson, and R. R. Bovbjerg, "Obtaining Providers' 'Buy-In' and Establishing Effective Means of Information Exchange Will Be Critical to HITECH's Success," *Health Affairs* 31, no. 3 (2012): 514-526, at 519.
16. J. Ahern, A. Hubbard, and S. Galea, "Estimating the Effects of Potential Public Health Interventions on Population Disease Burden: A Step-by-Step Illustration of Causal Inference Methods," *American Journal of Epidemiology* 169, no. 9 (2009): 1140-1147; see Cousens et al., *supra* note 4.
17. D. Faigman, J. Blumenthal, E. K. Cheng, J. L. Mnookin, E. E. Murphy, and J. Sanders, *Modern Scientific Evidence: The Law and Science of Expert Testimony* (Minnesota: Thomson Reuters/West, 2011): at §5:16, pp, 281-282.
18. S. Greenland, "Quantifying Biases in Causal Models: Classical Confounding vs. Collider-Stratification Bias," *Epidemiology* 14, no. 3 (2003): 300-306, at 306.
19. See Beasley et al., *supra* note 13, at 747; Faigman et al., *supra* note 17, at §5:10, at 277; G. P. Hammer, J. B. du Prel, and M. Blettner, "Avoiding Bias in Observational Studies," *Deutsches* Ärzteblatt *International* 106, no. 41 (2009): 664-668, at 665.
20. K. Turner and L. Ferland, "State Electronic Disease Surveillance Systems – United States, 2007 and 2010," *Morbidity and Mortality Weekly Report* 60, no. 41 (2011): 1421-1423, at 1421; H. Rolka, D. W. Walker, R. English, M. Katzoff, G. Scogin, and E. Neuhaus, "Analytical Challenges for Emerging Public Health Surveillance," *Morbidity and Mortality Weekly Report* 61, Supp. (2012): 35-39, at 36.
21. S. Sachdeva and S. Bhalla, "Semantic Interoperability in Standardized Electronic Health Record Databases," *Association for Computing Machinery Journal of Data and Information Quality* 3, no. 1 (2012): 1-1:37, at 1:5.
22. Optum, *A CIO Survey of HIT Adoption Trends*, An Optum Institute Survey Brief (2012), *available at* <http://institute.optum.com/research/featured-publications/cio-survey-of-hit-adoptiontrends/~/media/OptumInstitute/Page_Elements/Articles/OPTUM_CIO_HIT_Survey_Feb2012.pdf> (last visited January 9, 2013).
23. M. Klompas, J. McVetta, R. Lazarus, E. Eggleston, G. Haney, B. A. Kruskal, W. K. Yih, P. Daly, P. Oppendisano, B. Beagan, M. Lee, C. Kirby, D. Heisey-Grove, A. DeMaria, and R. Platt, "Integrating Clinical Practice and Public Health Surveillance Using Electronic Medical Record Systems," *American Journal of Preventive Medicine* 42, no. 6, Supp. 2 (2012): S154-S162. The ESP platform "automatically execute[s] complex disease-detection algorithms to provide meaningful surveillance without requiring clinicians to manually parse potential cases." *Id.*, at S154.
24. M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A Pragmatic Framework for Single-Site and Multisite Data Quality Assessment in Electronic Health Record-Based Clinical Research," *Medical Care* 50, Supp. (2012): S21-S29.
25. J. Pearl, *Causality*, 2d ed. (New York: Cambridge University Press, 2009): at 65-68; T. J. VanderWeele and N. C. Staudt, "Causal Diagrams for Empirical Legal Research: Methodology for Identifying Causation, Avoiding Bias, and Interpreting Results," *Law, Probability and Risk* 10, no. 4 (2011): 329-354.