Commentary: Electronic Health Records for Comparative Effectiveness Research

Russell E. Glasgow, PhD

The papers in this issue by Desai et al[1] on use of electronic health records (EHRs) to form disease registries and Kahn et al[2] regarding quality control in the EHR both discuss important and timely issues. Use of frameworks, strategies, and recommendations provided in these articles will go a long way toward improving the quality of EHR data for purposes ranging from local quality improvement to comparative studies and national population-based research.

This is clearly a case where "the devil is in the details," as is well documented in examples provided in these articles. The data quality and checking issues and multiple decisions that need to be made in creating registries are paramount, because electronic data are not inherently better or worse than paper records. As often happens, debates have often focused on the overall advantages and disadvantages of EHR data, instead of on the important subissues discussed in these papers, such as the purpose for which the data are intended, whether the project is single site or for a multi-site network, and whether priority should be given to sensitivity, specificity, or positive predictive value. These questions all have important implications for data cleaning and quality control, as well as interpretation of results. Attention to them is essential if EHR data are to be used to inform rapid learning health care systems[3] and to identify variations in care and outcomes.[4] If done correctly, EHR data have the potential to dramatically improve the quality of local care and to fuel much needed cross-site comparative effectiveness research.

The discussion of lessons learned in both papers is helpful, and at times sobering, such as having to drop an entire site from a network project due to ongoing quality control problems. Those not experienced in working with EHR data are likely not aware of all the issues involved in producing high-quality electronic data. Kahn et al[2] also make a seldom recognized but important point that it is virtually impossible and cost prohibitive to conduct comprehensive data checking and cleaning on all data and that research teams need to prioritize and decide upon which data elements are critical for a given question.

The issues addressed by Kahn et al[2] and Desai et al[1] are central to improving quality and in forming registries from data that are currently in EHRs. To produce a truly disruptive innovation in health care research, however, it will also be necessary to address issues regarding data that are *not* currently in most EHRs, and if they are, are not collected in any standardized way. Given space limitations, I briefly discuss 2 such types of data: patient-reported data and social environmental data. It is ironic that with all the focus on patient-centered comparative effectiveness research (http://www.pcori.org) and the patient-centered medical home,[5] the one type of data that are not collected routinely or in a standardized manner are patient-reported data.[6]

To operationalize personalized or precision medicine, it is essential to have data from the patient as part of the information base for decision making and tailoring. Patients are key and often the only feasible source of data on important health determinants such as health behaviors, mental health, and psychosocial issues.[6] In my opinion,

From the Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD.

The opinions expressed are those of the author and do not necessarily represent the opinion of the NIH.

The author declares no conflict of interest.

Reprints: Russell E. Glasgow, PhD, Implementation Science, Division of Cancer Control and Population Sciences, National Cancer Institute, 1630 Executive Blvd., Rockville, MD 20852. E-mail: glasgowre@mail.nih.gov.

ISSN: 0025-7079/12/5007-0S19

greater priority should be given to gathering these patient-reported data in standardized ways, as biometric and laboratory data are. Another essential type of patient-reported data not found consistently completed for most patients in EHRs are information on patient characteristics such as race, ethnicity, income, health literacy, and patient preferences. It is not possible to reduce health disparities if a system does not know which patients are members of different groups.

When a person's zip code predicts as much or more about their health status than his or her genetic code,[7] it is logical to think that information on his or her social and physical environment should also be part of the health record. Although not widely used today, the virtual explosion in social environmental data available through the Open Government Act and the corresponding wide array of community health indicator and geocoded data offer tremendous potential for tailoring interventions and understanding health care and health outcomes.

Although there are still some who think that a randomized controlled trial provides the best answer to all health and health care questions, many are coming to the conclusion that for many practical questions, the availability of real-world, close to real-time EHR data on hundreds of thousands or millions of real-world patients receiving care in real-world settings by typical staff are incredibly useful and possibly even preferred sources of data for applied questions[8]—assuming that the data can be trusted.

These articles by Desai et al[1] and Kahn et al[2] are important, because they discuss key issues in enhancing EHR data quality and provide useful and generalizable procedures to create disease registries and guide quality control efforts. The old adage of "garbage in, garbage out," applies to EHR data and most other types of health services research data. In closing, it is appropriate to also note that like all data sources, EHR data are not the answer to all issues. Although high-quality EHR data can be used for many purposes, in situations such as that described by Desai et al[1] where one has almost 50% disenrollment in a longitudinal cohort, these are not the optimal data source for such longitudinal questions.

REFERENCES

1. Desai J, Wu P, Nichols G, et al. Diabetes and Asthma Case Identification, Validation, and Representativeness When Using Electronic Health Data to Construct Registries for Comparative Effectiveness and Epidemiologic Research. *Med Care*. 2012;50(suppl 1):S30–S35.
2. Kahn MG, Raebel MA, Glanz JM, et al. A *Pragmatic Framework for Single and Multi-site Data Quality Assessment in Electronic Health Record-based Clinical Research. Med Care*. 2012;50(suppl 1):S21–S29.
3. Etheredge LM. A rapid-learning health system: What would a rapid-learning health system look like, and how might we get there? Health Affairs, Web Exclusive Collection, w107-w118. 2007 26 Jan [Epub ahead of print]. doi:10.1377/hlthaff.26.2.w107.
4. Wennberg JE. *Tracking Medicine: A Researcher's Quest to Understand Health Care*. New York: Oxford University Press; 2010.
5. Nutting PA, Crabtree BF, Miller WL, et al. Transforming physician practices to patient-centered medical homes: lessons from the national demonstration project. *Health Aff (Millwood)*. 2011;30:439–445.
6. Glasgow RE, Kaplan RM, Ockene JK, et al. The Need for Practical Patient-report Measures of Health Behaviors and Psychosocial Issues in Electronic Health Records. *Health Aff (Millwood)*. 2012;31: 3497–3504.
7. Marks JS. Why your zip code may be more important to your health than your genetic code. Huffington Post, April 23, 2009. http://www.huffingtonpost.com/james-s-marks/why-your-zip-code-may-be_b_190650.html. Accessed November 16, 2011.
8. Kessler R, Glasgow RE. A proposal to speed translation of healthcare research into practice: dramatic change is needed. *Am J Prev Med*. 2011; 40:637–644.