EDITORIAL AND COMMENT

Medicine Based Upon Data

Charles Safran, MD, MS

Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Brookline, MA, USA.

J Gen Intern Med 28(12):1545–6
DOI: 10.1007/s11606-013-2549-3

In this issue of *JGIM*, Shirts and colleagues[1] demonstrate how information from an institution's electronic health record can be used to tailor clinical decision support for a particular patient. Using the example of celiac disease, they highlight the importance of local context when aggregating clinical data. Not only does the sensitivity and specificity of a diagnostic procedure vary with the institution as compared to published averages, but these parameters also vary within the institution according to who is ordering the procedure.

The dream of evidence-based medicine 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 small fraction of clinical situations. Moreover, the inclusion and exclusion criteria mean that rarely is the evidence generated by RCTs strictly about "patients like my patient."

If we cannot always turn towards the literature for evidence, will humongous databases of routinely collected clinical data be an acceptable alternative?[2] The answer as shown by Shirts and colleagues[1] in this issue of *JGIM* is a qualified yes with some caveats. Important concerns involve the quality of the data, the limitations of methods used for analysis, and the changing nature of medical practice.

Clinical data are messy. Data are missing and may be inaccurate. Unlike data prospectively collected as part of a research protocol, there is no quality control, normalization, or regularity of the data. Much of the data we capture in electronic health records is unstructured clinical narrative and hence not easily used for analysis. Structured data are the most amenable for analysis. Almost all structured data are collected and stored for some reason. Some data, such as laboratory results, are automatically generated. Other data, such as conditions on a problem list, are only generated when a clinician chooses to enter the data. Historically, we collected information in the clinical setting to facilitate billing for services. More recently some of our data collection, such as documentation of smoking status, is driven by "meaningful use" reimbursement. Thus, routinely collected data come with inherent biases. Another source of bias is selection of patients by clinicians for specific interventions. Shirts and colleagues[1] revealed this type of bias by showing that patients of gastroenterologists were more likely to have positive biopsies for celiac disease than were patients of generalists. Lastly, clinical data are time-oriented and time-sensitive. Time relationships in clinical databases are not well represented. For instance, the time of initial diagnosis of diabetes might not be collected as part of a problem list, but would be essential for understanding some aspects of comparative effectiveness of alternative therapies. Clinicians might have tried tincture of time as a diagnostic intervention or the patient might have just delayed follow-up. This type of context, which is essential for the interpretation of clinical data, is often missing in the databases that are available for clinical research.

Because of these many types of biases and limitations, routinely collected clinical data present analytic challenges for the clinician hoping to use real world evidence in direct patient care. Shirts and colleagues have used a "simplified near-neighbor classification" to alleviate some of the concerns relating to bias[1] by comparing the closeness of a patient to those with a known outcome in a multi-dimensional space where each dimension is a confounding variable. While traditional clinical research is hypothesis driven, large data sets invite exploration and multiple comparisons. While some see potential gold in the mining of data, others are concerned that false associations will only yield fool's gold.

Our understanding of disease and hence treatment continues to evolve. When we first started building a large clinical data repository and making it available for clinical research in the 1980's[3] there were only four types of non-Hodgkin's lymphoma. By 2008, the World Health Organization's classification of lymphomas listed over 30 B-cell malignancies alone![4] The implication of genomics is that all common complex conditions are really multiple rare diseases. Welcome to the future world of personalized medicine and its implications for finding data about "patients like my patients." Although Shirts suggests that local data may provide better evidence for decision-making than a meta-analysis of all available published information,

Published online July 10, 2013

there will never be enough local data when we consider all the biologic signals we will be able to collect from each patient. Moreover, most institutions are not nearly large enough to gather enough data to duplicate Shirts' study.[1]

The imperative to combine data across our collective clinical experience remains compelling. Local and regional efforts to form health information exchanges (HIE) are a step in the right direction, although the lack of data sharing arrangements and the lack of true granularity of the data collected by these HIEs will limit our ability to provide real-time decision support from these data sources. Perhaps we should ask our citizens to contribute a full copy of their health data to a trusted public utility so that with a truly humongous clinical database, we might be able to practice medicine based on data.[5]

Corresponding Author: *Charles Safran, MD, MS; Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, 1330 Beacon Street Suite 400, Brookline, MA 02446, USA (e-mail: csafran@bidmc.harvard.edu).*

REFERENCES

1. **Shirts BH, Bennett ST, Brian RJ.** Using patients like my patient for clinical decision support: institution-specific probability of celiac disease diagnosis using simplified near-neighbor classification. J Gen Intern Med. doi:10.1007/s11606-013-2443-z.
2. **McDonald CJ, Hui SL.** The analysis of humongous databases: problems and promises. Stat Med. 1991;10:511–8.
3. **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.
4. **Jaffe ES.** The 2008 WHO classification of lymphomas: implications for clinical practice and translational research. Hematology. 2009;2009:523–31.
5. **Knottnerus JA, Dinant GJ.** Medicine based evidence, a prerequisite for evidence-based medicine. BJM. 1997;315:1109–10.