Invited Commentary

Christopher G. Chute*

* Correspondence to Dr. Christopher G. Chute, Department of Health Sciences Research, Mayo Clinic College of Medicine, 200 First Street SW, Rochester, MN 55905 (e-mail: chute@mayo.edu).

Initially submitted October 7, 2013; accepted for publication October 17, 2013.

Historically, clinical epidemiologic research has been constrained by the costs and time associated with manually identifying cases and abstracting clinical data. In this issue, Carrell et al. (Am J Epidemiol. 2014;179(6);749–758) report on their impressive success using natural language processing techniques to correctly identify cases of cancer recurrence among women with previous breast cancer. They report a 10-fold decrease in the need for chart abstraction, though with an 8% loss in case detection. This commentary outlines some recent history associated with the development of “high-throughput clinical phenotyping” of electronic health records and speculates on the impact such computational capabilities may have for observational research and patient consent.

Clinical case retrieval; electronic medical records; high-throughput clinical phenotyping; natural language processing

Abbreviations: EHR, electronic health record; SHARP, Strategic Health IT [information technology] Advanced Research Projects.

The availability of data has historically been critical to the practice of epidemiology. Although most practitioners of observational research are exquisitely trained in inferencing methods, ranging from simple tabulation to sophisticated machine learning methods, these methods are for naught in the absence of data to drop into the analytical mill. Generations of epidemiology students have honed their parameter-estimating skills on fastidiously collected, thoughtfully curated, and well-guarded data sets that would often define the competitive quality of a faculty research program, department, or even an entire graduate school. No data, no study, no how.

The June 2012 Data and Informatics Working Group Report to the National Institutes of Health Director cites an epochal transition from data generation as a rate-limiting step for biomedical science to “data management, communication, and interpretation” ((1), p. 5).” Indeed, Dr. Francis Collins has asserted that “The future of biomedical research depends upon our ability to support a research ecosystem that leverages the flood of biomedical data . . .” (2), reinforcing the notion that data, including clinical phenotyping data, have become not only plentiful but nearly overwhelming in volume, velocity, and variety (the 3 Vs of Big Data (3)).

In this issue of the Journal, Carrell et al. (4) publish an exemplar that may represent the practical future of clinical epidemiology. Using clinical features and parameters efficiently extracted from clinical notes, pathology reports, and other narratives, the authors compellingly demonstrate the value of computational methods to prescreen potential cases and controls of breast cancer recurrence from hospital records. Specifically, they use natural language processing techniques to parse these textual sources among women with previous breast cancer, identify terms and elements that suggest disease recurrence, and optimize an algorithm across overlapping 30-day time windows to maximize precision and recall using a “gold standard” training set. They then execute this algorithm on a separate testing set, resulting in a 10-fold reduction in the number of charts requiring manual review, at the arguably modest cost of potentially missing 8% of women with a true recurrence. They comment thoughtfully on how the balance between precision and recall can be “tuned” by parameters in the case detection algorithms.

What Carrell et al. (4) are demonstrating is an example of what the medical informatics community calls “high-throughput clinical phenotyping” of electronic health records.
(EHRs). Indeed, an entire special issue of the Journal of the American Medical Informatics Association dedicated to this topic is about to appear; the editor received more than 60 high-quality manuscript submissions in response to the call for papers, an unusually large number in the relatively small community of medical informatics. The phenotyping moniker is badly overloaded semantically but has come to represent cohort identification for research (as in the current issue), clinical trial eligibility, numerators or denominators of quality metrics, or even the cohort of patients for whom a specific clinical decision support reminder should “ﬁre.” The term was inspired by the recognition that, although we can easily and inexpensively genotype large patient cohorts quite rapidly, the rate-limiting step in genotype-to-phenotype research—for example, genome wide-association studies—remains the ability to characterize patients with respect to their clinical phenotypes.

The National Institute of Health’s National Human Genome Research Institute sought to address this phenotyping gap by forming the Electronic Medical Records and Genomics (eMERGE) Network in 2007 (5, 6). In the course of conducting genome-wide association studies across multiple organizations, the need arose to develop and deploy phenotyping algorithms that would provide high positive predictive value when executing over heterogeneous EHRs at different academic medical centers. This was demonstrated to work for specific diseases, such as peripheral arterial disease (7) and type 2 diabetes mellitus (8), and it was also validated for many disease phenotype algorithms across the eMERGE Network (9, 10). Indeed, an early peripheral arterial disease algorithm was based primarily on radiology notes using natural language processing (11), using essentially the same Mayo Clinic–developed open-source natural language processing software (12) as that used by Carrell et al. (4). For the genome-wide association studies use case, where we could optimize for precision and have less concern about recall, these fully algorithmic phenotyping methods were used over millions of patient records from many academic medical centers, yielding—without any manual review—thousands of cases that now form the basis of a large corpus of scientiﬁc papers (across multiple diseases) published by network members.

Obviously, fully automated, algorithmic phenotyping across medical records is neither easy nor perfect. Some algorithms entail substantial complexity, deﬁning several hundred decision elements, such as inclusion and exclusion criteria, in machinable form (13). We have also demonstrated that these algorithms can work poorly for clinical records that are fragmented over many clinical providers, such as for patients seen by a primary care provider and a tertiary academic medical center (14). Not surprisingly, we also demonstrated that the longitudinal duration of patient data is correlated with retrieval performance: the longer the patient history, the more accurate the phenotyping behavior (15).

In the past 5 years, the adoption of “meaningful use” (16) policies as part of the federal Health Information Technology for Economic and Clinical Health Act has caused the prevalence of EHRs to rise from less than 10% to nearly 70%; EHRs have arrived, and with them, a vast wealth of clinical data readily retrievable using computation methods. Using American Recovery and Reinvestment Act (ARRA) of 2009 funding, the Ofﬁce of the National Coordinator for Health Information Technology sponsored 4 large grants to facilitate access to data within EHRs, called the Strategic Health IT Advanced Research Projects (SHARP). One of them—called SHARPn, where the “n” is for normalization—was awarded to the Mayo Clinic (Rochester, Minnesota) and its consortium of academic and commercial partners and focused on the secondary use of patient records (http://sharpn.org). SHARPn is developing a suite of open-source tools for clinical data normalization, including natural language processing (which SHARPn considers a kind of data normalization from unstructured documents), and broadly executable clinical phenotyping algorithms (17, 18). These, and similar software resources for normalized data retrieval and case ﬁnding from EHRs, can be expected to transform the practicality of high-throughput clinical phenotyping from the experimental to common practice throughout biomedical and translational research programs.

The implications for observational research on patient outcomes, and clinical epidemiology in particular, should be obvious. The costs in time and resources for algorithmic case and control identiﬁcation, electronic harvesting of normalized data into study data sets and registries, and linkage for the discovery of potential clinical outcome events across a myriad of clinical providers in a study region will all be dramatically reduced compared with the historical model of human chart abstraction. Although the long-sought vision of being able to do an entire epidemiologic study “in an afternoon” may remain elusive, the proportion of time and resources that can be focused on data integrity and consistency measures, case validation, and thoughtful analyses can be correspondingly increased.

It would be remiss in any commentary on EHR data access to overlook the increasingly important issues of privacy, conﬁdentiality, data security, and consent. Perhaps consent is the central issue. Historically, patients have consented to be in a speciﬁc study conducted by a speciﬁc research team. Now, however, patients and providers will be confronted with whether and how to consent for broad, future, and potentially unspeciﬁed uses of clinical information in the cause of research. Although deidentiﬁcation and “safe harbor” HIPAA provisions provide some protections, the opportunity for practical cross-institutional data linkage to avoid the data fragmentation problem that is inevitable when care is received from multiple providers (14) bears careful consideration as to when deidentiﬁcation should occur in the data management process, and by whom. Regardless, scarcity of data may now be a historical condition that most clinical epidemiologists are unlikely to confront in the near future.

ACKNOWLEDGMENTS

Author afﬁliation: Biomedical Informatics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota (Christopher G. Chute).

Conﬂict of interest: none declared.
REFERENCES