Response to Invited Commentary

Carrell et al. Respond to “Observational Research and the EHR”

David S. Carrell*, Scott Halgrim, Diem-Thy Tran, Diana S.M. Buist, Jessica Chubak, Wendy W. Chapman, and Guergana Savova

* Correspondence to Dr. David Carrell, Group Health Research Institute, Metropolitan Park East, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101 (e-mail: carrell.d@ghc.org).

Initially submitted December 9, 2013; accepted for publication December 20, 2013.

We appreciate Dr. Chute’s thoughtful comments and perspectives on the broader landscape of clinical informatics (1). Three points he makes regarding clinical natural language processing (NLP) of electronic health record (EHR) data have important implications for future research.

First, the availability of clinical text for secondary use is a critical first step toward realizing its potential. As Dr. Chute notes, we are witnessing a transition from a past constrained by data to a future that may be overwhelmed by it. Yet in most institutions, researcher access to EHR text lags far behind access to structured data. Barriers to accessing clinical text tend to be less about technical challenges than about perceptions—we would argue, misperceptions—of risks to patient privacy once text is made available outside the EHR. Correcting potential misperceptions is a critical first step in improving access. Difficulties in accessing EHR text across multiple institutions is an additional challenge. Health information exchange addresses this issue from the perspective of care delivery but does not necessarily address aggregating the clinical narrative for research. Longitudinally incomplete records, as Dr. Chute points out, retard algorithm performance. Our own work illustrates this. Four of our NLP algorithm’s 5 case identification failures were due to the absence of electronic versions of relevant chart notes. Overcoming access challenges would benefit observational investigations from small to large. Consider, for example, the potential impact on postmarket drug safety surveillance through the Food and Drug Administration’s Sentinel Initiative (2, 3) if structured data from the more than 130 million covered lives from 18 organizations could be augmented with text-based EHR information.

Second, advancing clinical NLP methods will benefit from open source, sharable resources and software. Dr. Chute mentions the important contributions made by the Strategic Health IT [information technology] Advanced Research Projects (SHARP). Like others, we have benefitted from SHARP-produced tools, including software modules that detect when concepts of interest are qualified by negation or uncertainty—essential details for interpreting cancer screening and diagnosis narratives. Indeed, in the spirit of advancing portability and reproducibility, we have contributed to the Apache cTAKES (4) NLP system. Also essential are sharable samples of clinical text that have been annotated for syntactic and semantic content. Availability of such annotated text in the general domain since the 1990s (5) propelled NLP far beyond the current state of clinical NLP. Achieving parity in the clinical domain, particularly in the application of powerful machine learning approaches, will require additional sharable resources. To accelerate the broad use of clinical NLP technologies, developers should also strive to make their systems more easily portable, including to settings where informatics expertise is limited, thereby extending their potential impact.

Third, protecting patient privacy is critically important when using EHR data for research. We agree with Dr. Chute that text deidentification will play an important role in facilitating access while protecting patient privacy. Methods to address the trace amounts of residual identifiers that automated deidentification tools overlook are needed (6). That said, machine processing has inherent advantages over manual abstraction. Though automated systems may “touch” far greater quantities of clinical text, they entail substantially less exposure of patient information than manual abstraction because relatively little human review is needed. Institutional review boards appreciate this. Indeed, the Group Health (Seattle, Washington) institutional review board welcomed the adoption of NLP methods and the construction of a research database for EHR text because they were convinced this would reduce threats to patient privacy. We also point out that, once text-based information has been locally extracted, it can be normalized to standardized coding schemes, free of
direct patient identifiers, allowing it to be more easily sharable.

ACKNOWLEDGMENTS

Author affiliations: Group Health Research Institute, Seattle, Washington (David S. Carrell, Scott Halgrim, Diem-Thy Tran, Diana S.M. Buist, Jessica Chubak); Division of Biomedical Informatics, Department of Medicine, University of California San Diego, San Diego, California (Wendy W. Chapman); Informatics Program, Boston Children’s Hospital, Boston, Massachusetts (Guergana Savova); Department of Pediatrics, Harvard Medical School, Harvard University, Boston, Massachusetts (Guergana Savova); and Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington (Jessica Chubak).

Conflict of interest: Guergana K. Savova is on the Advisory Board of Wired Informatics, LLC, which provides services and products for clinical natural language processing applications.

REFERENCES