Natural Language Processing and the Promise of Big Data
Small Step Forward, but Many Miles to Go

Thomas M. Maddox, MD, MSc; Michael A. Matheny, MD, MS, MPH

The promise of big data has captured healthcare’s imagination. Although the term lacks a consensus definition, it generally refers to electronic health data sets characterized by the 3 Vs: volume, variety, and velocity.1,2 Volume refers to the sheer amount of healthcare data currently generated by clinical operations, administration, and patients themselves. By one estimate, ≈ 25,000 petabytes of healthcare data will be available by 2020—an amount that could fill 500 billion file cabinets.2 Variety refers to the wide range of healthcare data formats. For example, electronic health records (EHRs) contain both structured and unstructured (or free-text) data, diagnostic images come in a variety of multimedia formats, and patient data are generated from wearables, mobile devices, medical devices, and social media—each with its own format. Velocity refers to the rapidity with which new data are generated, and thus the speed at which it needs incorporation into data sets and analyses to provide real-time insights into patient care.

The potential of such data is enormous. Insights from big data could fuel innovation and improvement in clinical operations, research and development, and public health.1 However, the potential of big data to realize these lofty aspirations is matched by the challenge of organizing, analyzing, and generating actionable insights from it.

One of the biggest challenges in realizing the potential of big data is in abstracting it. With the passage of the HITECH (The Health Information Technology for Economic and Clinical Health) Act in 2009, the adoption of EHRs in clinical practice has accelerated, and now over half of office-based clinical settings have used keyword searching and other techniques to facilitate quality and safety monitoring in a variety of healthcare settings.7,8 However, the structured data rather than a standalone tool. In more clinical settings, the shotgun NLP tools do not perform well enough for focused clinical tasks like real-time surveillance, quality profiling, and quality improvement initiatives, and the focused NLP tools tend to lose performance in clinical environments outside of their development frame. Scaling NLP text processing to handle system-wide processing of large volumes of generated text data has also been a challenge. As a result, NLP use in clinical operations has been limited.

Wasfy et al9 add to these initial NLP attempts to optimize healthcare delivery in this issue of Circulation: Cardiovascular Quality and Outcomes. In their study, they concentrate on patients undergoing percutaneous coronary intervention (PCI) and their risk for hospital readmission in the 30 days after the procedure. Improving our ability to predict those at risk for readmission and the modifiable reasons behind it...
is an important goal. Substantial literature exists to suggest that some readmissions are preventable with changes in care delivery.\textsuperscript{10–13} In addition, improvement in preventable readmissions is an increasingly important reputational and financial priority for health systems.\textsuperscript{14} This emphasis on readmission reductions is an increasingly important reputational and financial priority for health systems.\textsuperscript{14} This emphasis on readmission reduction has recently turned to post-PCI patients. Currently, risk-standardized all-cause unplanned post-PCI readmission rates range from 8.6% to 16.8%.\textsuperscript{15} Successful efforts to reduce these readmissions will require accurate readmission prediction models.

Current prediction models for PCI readmission are derived from structured clinical and claims data. Accordingly, the researchers used NLP to incorporate unstructured EHR data in an attempt to improve the discrimination of their current models. They selected candidate predictor variables from previous readmission studies, including 8 variables that were only available in an unstructured format, and then used a locally developed NLP tool to extract both structured and unstructured data. They found that 3 variables—number of emergency department visits, anticoagulation, and a provider-recorded assessment of patient anxiety—were all significantly associated with readmission and increased discrimination compared with the prior model. The authors’ work provides important validation of the value of incorporating NLP into prediction efforts and is novel proof-of-concept work. The most direct application of these findings is reducing the burden of manual chart review by reliable filtering of irrelevant documents using NLP—an important and costly function in quality improvement. In addition, the 3 novel variables identified—number of emergency department visits, anticoagulation, and anxiety—have not been previously incorporated into post-PCI readmission prediction models and may improve their discrimination and calibration. The anxiety variable is particularly compelling because it may relate to a wide variety of behavioral health and social issues that both drive readmission and are modifiable. The anxiety definition used by the authors was broad, so more work needs to be done to understand the association more clearly, identify potentially preventable and modifiable components, and to inform possible interventions.

The contributions that this study makes to both NLP and additional variables that drive post-PCI readmissions should be taken in the context of several caveats about the study design and limitations of the NLP approach used. First, the study design is not optimal for evaluating the true ability of the NLP tool to meaningfully improve readmission prediction with unstructured data. Unstructured data, when incorporated into a pre-existing model comprised structured data only, should significantly improve prediction discrimination and calibration, relative to the structured data only model. The investigators’ decision to use a case–control methodology, matching on the structured data elements of the pre-existing model, prevents direct comparison between models. Second, the readmission risk factors identified are more correctly attributed to the risk of readmission to the same hospital rather than readmission to any hospital because this work only looked at readmissions to the index hospital, and about 33% of study patients were readmitted to hospitals different than the one where the PCI was performed. Although the authors partially addressed this situation with a sensitivity analysis, differences in these populations may affect the relevance of the model to those readmitted to hospitals different than the index one. Second, the low occurrence of some of the candidate variables—homelessness, cirrhosis, and syncope or presyncope—may explain their lack of significance in the prediction model. Prior literature suggests that these conditions are associated with readmission in non-PCI populations, so more exploration may be warranted for these conditions. Finally, Soothsayer (Boston, MA), the specific research NLP tool used in this analysis, uses a regular expression parser technique for NLP. This approach does not allow for important features available in more robust NLP tools, such as accounting for misspellings, synonyms, disambiguation of expressions by parts of speech and proximity words, negation, and temporality. Thus, Soothsayer’s ability to conduct automated surveillance and quality profiling may be limited. In addition, this may limit its generalizability outside of the investigators’ institution because of systematic differences in how healthcare systems generate clinical care documentation.

Several next steps are needed to drive the usefulness of NLP in improving insights into risk factors for PCI readmission and ultimately for a broader group of health outcomes. First, understanding and optimizing the ability of NLP tools to accurately identify important data elements, without the need for manual chart verification, are critical to efficiently abstract unstructured data and make the process scalable. Second, using a mixture of deductive, a priori clinical variable selection—informing prior studies and clinical reasoning—and inductive, data-driven variable selection—driven by patterns seen in the EHR data—are needed to maximize the information available to predict PCI readmission. Finally, testing the performance of Soothsayer, and other NLP tools, in a variety of data sets and clinical settings will be necessary to both understand drivers of readmissions across settings and inform the tool’s use for other clinically valuable tasks. This testing and optimization of NLP tools need to account for differences not only across clinical settings but also across providers and time. Provider training, temperament, and speciality all drive differences in documentation style and nomenclature, and effective NLP tools will need to account for all of these variables to be effective. Similarly, documentation style can change over time—sometimes directly in response to the knowledge that NLP tools will be reading and abstracting the data—and periodic testing of the tools will be needed to guard against a loss of accuracy over time.

The amount of data in health care is immense and growing. Maximizing our use of this data is fundamental to the creation of learning healthcare systems and its goal of best care at lower cost.\textsuperscript{16} Many miles remain in our journey toward this goal, and continued innovation and validation of NLP and other tools in extracting meaningful insight from unstructured data are critical. Wasyf et al’ provide a small, but important, step forward.

Disclosures

None.
References


Key Words: Editorials database electronic health records natural language processing percutaneous coronary intervention
Natural Language Processing and the Promise of Big Data: Small Step Forward, but Many Miles to Go
Thomas M. Maddox and Michael A. Matheny

Circ Cardiovasc Qual Outcomes. 2015;8:463-465; originally published online August 18, 2015;
doi: 10.1161/CIRCOUTCOMES.115.002125
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272
Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/8/5/463

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation: Cardiovascular Quality and Outcomes can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation: Cardiovascular Quality and Outcomes is online at:
http://circoutcomes.ahajournals.org//subscriptions/