Editorial: Managing unstructured and structured big data in healthcare system: generating valid real-world evidence by natural language processing

Kevin Lu*

Department of Clinical Pharmacy and Outcomes Sciences,
College of Pharmacy,
University of South Carolina,
Columbia SC 29208, USA
Email: lu32@email.sc.edu
*Corresponding author

Minghui Li

Department of Clinical Pharmacy and Translational Science,
University of Tennessee Health Science Center,
Memphis TN 38163, USA
Email: mli54@uthsc.edu

Jun Wu

Presbyterian College School of Pharmacy,
Clinton, SC 29325, USA
Email: jwu@presby.edu

As we continue to move into the digital health era, vast quantities of healthcare data are being accumulated and mad available for research and treatment decisions. Faced with the challenges of the volume, velocity, variety, and veracity (4Vs) of healthcare big data, healthcare systems and researchers need to adopt new technologies for collecting, storing, and analysing large-scale data to generate better real-world evidence for more informed decision-making. Clinical 'big data' stored in electronic health record (EHR) systems are generally formatted as either structured or non-structured datasets (Weber et al., 2014; U.S. Centers for Medicare & Medicaid Services, https://www.cms.gov/medicare/e-health/ehealthrecords/index.html).

Structured data refers to datasets containing variables (e.g., demographics, medication list, patient vitals, lifestyles and family history) in the same consistent format (Ruch, 2009). Structured datasets have been the primary format in the healthcare system in the past decades because they are easily tracked and readily processed by computers. However, structured data are limited in that they often follow fixed data models and value sets, which only allow pre-determined, limited values or formats. Unlike structured data, unstructured data are not arranged by pre-defined data models or schema and are not stored in a fixed record length format (Ruch, 2009). Unstructured data come from a variety of different formats, such as images, audios, videos, or unstructured texts.

Historically, researchers have focused on structured data and predictive analytics while ignoring unstructured data. However, unstructured data are the largest component of big data and account for 95% of available data while structured data form only a small portion of big data (Gandomi and Haider, 2015).

The most difficult challenge posed by unstructured data is that they are not uniform in formats, difficult to consolidate and standardise, often not a good fit for a mainstream relational database, and thus require more comprehensive analyses. Consequently, it is often difficult to generate useful real-world evidence with efficiency and scientific validity with unstructured data (e.g., unstructured clinical texts).

One possible solution to address the challenges of unstructured data with text is to use natural language processing (NLP) (Yim et al., 2016). NLP uses computational algorithms to understand human language (e.g., clinician-generated narrative text), and is an essential component of artificial intelligence (AI). NLP can be useful for abstracting information from unstructured data, thus allowing researchers to sift through overwhelming troves of free texts to find key information of interest (Doan et al., 2014).

This issue of the *International Journal of Computational Medicine and Healthcare* presents two studies by He et al. which examined the performance of NLP in structured data versus unstructured datasets. In the first paper, the authors used electronic health records (EHRs) from two large institutions and compared the differences between NPL and query methods to identify patients with metastatic melanoma. They reported that:

1 The NLP method of unstructured data analysis identified more patients with metastatic melanoma than structured data query methods (1,727 vs. 607 patients).

2 NLP had a statistically better sensitivity than structured query for patient identification (67% vs. 35%, $p < 0.05$) based on an external tumour registry.

Based on these findings, the authors concluded that NLP should be used to identify potential cancer study candidates with metastatic disease.

In the second study, He et al. examined the effectiveness and performance of using an NLP algorithm in unstructured data to identify important breast cancer biomarkers in patients with breast cancer. Specifically, they used nDepth, Regenstrief NLP platform to develop an NLP algorithm for identification of three biomarker status, namely estrogen receptors (ER), progesterone receptors (PRs), and human epidermal growth receptor factor 2 (HER2). The authors reported the performance of the NLP algorithms for extracting ER, PR, and HER2 receptor status ranging from 87.5% to 92.6% for sensitivity, 88.6% to 95.8% for specificity, 82.4% to 99.0% for positive predictive values (PPV), and 85.2% to 97.7% for negative predictive values (NPV). The authors concluded that NLP algorithms for unstructured data can be effective for the identification of important biomarkers in patients with breast cancer.

As correctly pointed out by the authors, limitations of these two studies include lack of chemotherapy data in the NLP algorithm, limited number of participating institutions (only two institutions), lack of inter-rater reliability of the chart reviewers, and potential selection bias. More importantly, both studies only focused on a single form of cancer and/or metastatic status. The literature documents that the performance of NLP could vary significantly for different cancer types or metastatic status (Yim et al., 2016; Spasić et al., 2014; Cheng et al., 2010). Despite these caveats, the two studies highlight the importance and feasibility of employing NLP techniques to enhance traditional structured data analysis methods.

Earlier this year, the Food and Drug Administration (FDA, https://www.sentinelinitiative.org/communications/publications/sentinel-system-five-year-strategy-2019-2023) released its five-year strategy for sentinel system (2019–2023). The Sentinel Initiative, launched in 2007 and fully implemented in 2016, is multi-site, privacy-preserving, curated data infrastructure and suite of analysis tools. It has now become one of FDA's premier platforms to generate real-world data and real-world evidence. Over the next five years (2019–2023), the FDA will expand its access to and use of EHRs by focusing on new technologies emerging from new data science disciplines, such as NLP. Specifically, the FDA (https://www.sentinelinitiative.org/communications/publications/sentinel-system-five-year-strategy-2019-2023) will "establish standards for NLP of unstructured data, including best practices for regulatory use." It is expected that incorporation of data derived from NLP will enable the sentinel system to identify valuable clinical and patient information in unstructured text in EHRs, or to identify previously undetected complex health outcomes. NLP will be a core feature of FDA's post-market safety surveillance armamentarium and a vital testbed for advanced technologies and approaches in the near future. In fact, the FDA has initiated an ongoing Validation of Anaphylaxis Using Machine Learning Project to develop a methodological framework for improved health outcome identification algorithms using machine learning and NLP techniques (FDA, https://www.sentinelinitiative.org/communications/publications/sentinel-system-five-year-strategy-2019-2023).

NLP of clinical text can and will offer a plethora of rich evidence from the real-world that should dramatically shift our current focus on structured data to unstructured data. The next step for NLP of clinical texts is to improve the efficiency and scientific validity by developing better tools (e.g., algorithms) to analyse unstructured big data in the healthcare system.

References

Cheng, L.T., Zheng, J., Savova, G.K. and Erickson, B.J. (2010) 'Discerning tumor status from unstructured MRI reports – completeness of information in existing reports and utility of automated natural language processing', *J. Digit. Imaging*, Vol. 23, No. 2, pp.119–132.

Doan, S., Conway, M. and Phuong, T.M. (2014) 'Natural language processing in biomedicine: a unified system architecture overview', *Methods Mol. Biol.*, Vol. 1168, pp.275–294, DOI: 10.1007/978-1-4939-0847-9_16.

FDA, *Sentinel System Five-year Strategy: 2019–2023* [online] https://www.sentinelinitiative.org/communications/publications/sentinel-system-five-year-strategy-2019-2023 (accessed 8 May 2019).

Gandomi, A. and Haider, M. (2015) 'Beyond the hype: big data concepts, methods, and analytics', *International Journal of Information Management*, Vol. 35, No. 2, pp.137–144.

Ruch, P. (2009) 'A medical informatics perspective on decision support: toward a unified research paradigm combining biological vs. clinical, empirical vs. legacy, and structured vs. unstructured data', *Yearb. Med. Inform.*, pp.96–98.

Spasić, I., Livsey, J., Keane, J.A. and Nenadić, G. (2014) 'Text mining of cancer-related information: review of current status and future directions', *Int. J. Med. Inform.*, Vol. 83, No. 9, pp.605–623.

U.S. Centers for Medicare & Medicaid Services, *Electronic Health Records* [online] https://www.cms.gov/medicare/e-health/ehealthrecords/index.html (accessed 5 September 2019).

Weber, G.M., Mandl, K.D. and Kohane, I.S. (2014) 'Finding the missing link for big biomedical data', *JAMA*, Vol. 311, No. 24, pp.2479–2480.

Yim, W.W., Yetisgen, M., Harris, W.P. and Kwan, S.W. (2016) 'Natural language processing in oncology: a review', *JAMA Oncol.*, Vol. 2, No. 6, pp.797–804.