Cooperative Oncology Groups. *Clin Cancer Res*. 2012;18(1):256-262. doi:10.1158/1078-0432.CCR-11-1633

35. Nasukawa T, Yi J. Sentiment analysis. In: Proceedings of the International Conference on Knowledge Capture—K-CAP '03. New York, New York: ACM Press; 2003:70.

36. Mullard A. Learning from exceptional drug responders. *Nat Rev Drug Discov*. 2014;13(6):401-402. doi:10.1038/nrd4338

37. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102. doi:10.1001/jama.2018.11100

38. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA*. 2018;320(11):1107-1108. doi:10.1001/jama.2018.11029

39. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22:1345-1359. doi:10.1109/TKDE.2009.191

40. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. *Cancer*. 2001;91(8)(suppl):1615-1635. doi:10.1002/1097-0142(20010415)91:8+<1615::AID-CNCR1175>3.0.CO;2-L

41. Ratwani RM, Savage E, Will A, et al. A usability and safety analysis of electronic health records: a multi-center study. *J Am Med Inform Assoc*. 2018;25(9):1197-1201. doi:10.1093/jamia/ocy088

42. Li X, Fireman BH, Curtis JR, et al. Privacy-protecting analytical methods using only aggregate-level information to conduct multivariable-adjusted analysis in distributed data networks. *Am J Epidemiol*. 2019;188(4):709-723. doi:10.1093/aje/kwy265

Invited Commentary

The Tipping Point for Deep Learning in Oncology

Andrew Daniel Trister, MD, PhD

Related article page 1421

Approaches to artificial intelligence have rapidly transformed many elements in our lives. Many have claimed that a new revolution fueled by artificial intelligence will bring even greater benefits and risks. Most of the common approaches leverage deep neural networks—algorithmic structures that process data through multiple layers of mathematical operations to result in a prediction or classification. These deep learning techniques require large amounts of well-annotated data and access to fast computing resources to ensure reasonable performance.

There are many promising applications of deep networks that are well positioned to affect diagnosis, research, and clinical decision-making.[1] Two limitations of deep learning techniques to drive this paradigm shift in clinical care relate to interpretability of the neural network prediction and generalizability of the results from a network to populations that were not included in the original training. To address these limitations, there are emerging best practices regarding assembling appropriate data sets; how those data are handled between training, testing, and validation; and methods to provide results that are directly understandable to humans (explainability).

In an article in *JAMA Oncology*, we see an example of the power of deep natural language processing to predict clinically relevant oncologic end points from radiologic reports that addresses some of these potential limitations of deep learning approaches.[2] Kehl et al[2] leveraged an existing set of electronic health record data collected for a precision medicine initiative at a single institution, Dana-Farber Cancer Institute, over the course of 5 years. The authors previously recognized the importance of a robust annotation framework of data collected in a clinical setting to inform precision medicine initiatives. They have described a novel structured framework for curation of clinical outcomes among patients with solid tumors using medical records data, PRISSMM, that ensures provenance and annotation from all data sources within the health system across pathology, radiology, signs and symptoms, molecular markers, and medical oncologist assessments toward clinical outcome assessment. They highlight that there are high upfront costs in this type of manual curation and that the implementation of algorithms to annotated records showed significant reduction in the time needed to process large numbers of records.

In their study, Kehl et al[2] restricted the training and testing of their algorithm to radiologic reports obtained from patients with a diagnosis of primary lung cancer. Reports were included regardless of histologic characteristics, stage, history of treatment, or modality of imaging included in the electronic health record. They took appropriate steps to separate the data into testing, training, and validation sets, and did so by randomly assigning patients and not individual reports into each group—an important step to ensure that information about a patient is not shared between phases of neural network development.

Despite reporting significant area under the receiver-operating characteristic curves for classifying a report along important clinical end points (existence of cancer, improvement, progression, and specific areas of concern for metastasis), the authors also incorporated local-interpretable model-agnostic explanation, which is a method that explores small changes to input data and the output of the model, giving some level of interpretability of the model.[3] The local-interpretable model-agnostic explanation algorithm highlights key words within specific reports that the model used to determine classifications. By highlighting elements of the data, the local-interpretable model-agnostic explanation provides the end user an opportunity to begin to evaluate how the algorithm made a specific determination. This additional layer of data provides a check against the black-box nature of the algorithm and should be standard in solutions for clinical decision support.

One limitation the authors did not mitigate in their approach is the generalizability of the model to other patients with lung cancers treated in settings outside Dana-Farber Cancer Institute. Although they used self-reported race to inform the discussion around how their model could be used in other settings, geographic and socioeconomic factors not included in their evaluation might show that the retrospective data included in their study are not representative of the mix of patients seen at other centers. Furthermore, despite Dana-

Farber Cancer Institute having a large faculty of radiologists interpreting the imaging studies, there is a high likelihood that the structure of the reports and vocabulary used are conserved across radiologists within Dana-Farber Cancer Institute with potential differences compared with radiologic reports sources from other centers.

A simple way to consider generalizability is whether there may be latent factors in the training data that would affect the performance of the model in another setting; in other words, the network can only know what it has already seen, and that bias in the construction of the training data set limits the performance when given other data.

One solution to decrease bias and improve generalizability is to make large data sets more widely available, both between institutions and openly among researchers. Such data sharing has been politically difficult and often infeasible owing to the sensitive and regulated nature of data generated within the health system. An alternative approach along these lines is to establish a consortium of institutions wherein trained networks are shared and performances are reported between member institutions, but no data are exchanged. By agreeing to evaluate models from other facilities, researchers at member institutions get a better sense of how well their models work and whether their approach could affect care elsewhere. A requirement for such a system to work is interoperability of the data fed to the model. Although we are not quite at the point of complete interoperable electronic health record data, new standards, such as fast health care interoperability resources, could make sharing models feasible.[4] Consortia could eventually agree to combine models to build a consensus network that could have advantages of increased generalizability at the cost of worse performance at each institution.

A third path toward improving generalizability could be to have a third party hold an established validation set that is broadly representative to compare the performance of individual trained models and also audit the performance in underrepresented groups. Although there are many such standard bearers across multiple agencies in government, this last path may prove to be impractical and has the potential to be abused. In addition, in other domains of deep learning research, there is increasing interest in training networks on synthetic data. In extreme cases, these synthetic data are generated by other neural networks. Such generative networks can build very large representative data sets, but there is a host of unknowns in the performance of such models in medical domains.[5]

Among the greatest hopes for artificial intelligence in medicine is the potential to both lower barriers to care and improve outcomes for large populations. Efforts that leverage the vast amount of data already digitized in the health system are reasonable first steps toward this promise. Just as our practice of oncology has been transformed by molecular techniques in the past decade, artificial intelligence will transform how we care for patients in the next decade. As both the generators of the data being in these solutions and ultimate end users of clinical neural networks, it is incumbent upon clinicians to help shape how we work together to improve the lives of patients.

ARTICLE INFORMATION

Author Affiliation: Department of Radiation Medicine, Oregon Health & Science University, Portland.

Corresponding Author: Andrew Daniel Trister, MD, PhD, Department of Radiation Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, L337, Portland, OR 97239 (trister@alumni.upenn.edu).

Published Online: July 25, 2019. doi:10.1001/jamaoncol.2019.1799

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z

2. Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports [published online July 25, 2019]. *JAMA Oncol*. doi:10.1001/jamaoncol.2019.1800

3. Tulio Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. Cornell University. http://arxiv.org/abs/1602.04938. Updated August 9, 2016. Accessed May 10, 2019.

4. HL7.org. Release 4. Welcome to FHIR. https://www.hl7.org/fhir/. Updated December 27, 2018. Accessed May 10, 2019.

5. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Cornell University. http://arxiv.org/abs/1809.07294. Updated March 5, 2019. Accessed May 10, 2019.