Practical guidance on artificial intelligence for health-care data

Advances in machine learning and artificial intelligence (AI) offer the potential to provide personalised care that is equal to or better than the performance of humans for several health-care tasks.[1] AI models are often powered by clinical data that are generated and managed via the medical system, for which the primary purpose of data collection is to support care, rather than facilitate subsequent analysis. Thus, the direct application of AI approaches to health care is associated with both challenges and opportunities.[2]

Many AI approaches use electronic health record (EHR) data, which document health-care delivery and operational needs, and which can be relevant to understanding patient health. EHR data are heterogeneous and are collected during treatment to improve each patient's health. Almost exclusively, EHR data are documented without consideration of the development of algorithms. Data can be collected from a wide range of sources, from high-frequency signals sampled every 0·001 seconds, to vital signs noted hourly, imaging or laboratory tests recorded when needed, notes written at care transitions, and static demographic data. Longitudinal clinical data, such as hospital records for all patients with a particular disease, are similarly heterogenous in data type, time scale, sampling rate, and reason for collection. Each data type is associated with its own challenges (figure). A glossary of technical terms is presented in the appendix (p 1).

High-frequency monitors record clinical signals (eg, oxygen saturation) with little human interaction, but they provide only a narrow view of patient state. These signals have frequent artifact corruption (eg, from sensors becoming dislodged), and must be aggregated, filtered, or discarded to remove artifacts. For example, electrocardiogram signals acquired in the USA must be filtered at 60 Hz to remove power grid electrical interference.

Imaging data, vital signs, laboratory tests, and other numerical measurements are ordered irregularly, and therefore they can produce biased data. Health-care workers might preferentially record data that are consistent with their understanding of patient state. For example, if a clinician suspects a particular diagnosis, he or she might record data that supports that diagnosis only. Additionally, clinicians often order tests related to the amount of variability they expect; therefore, the absolute time that a laboratory measurement is taken can be more predictive of patient health than the test value.[3] For example, health-care workers would only wake a patient at 0200 h to perform a blood test if the patient were very ill.

Narrative clinical notes are designed to provide a brief overview of the most important aspects of a patient's condition. However, standard AI tasks, such as word sense disambiguation (appendix p 1), are particularly

See Online for appendix

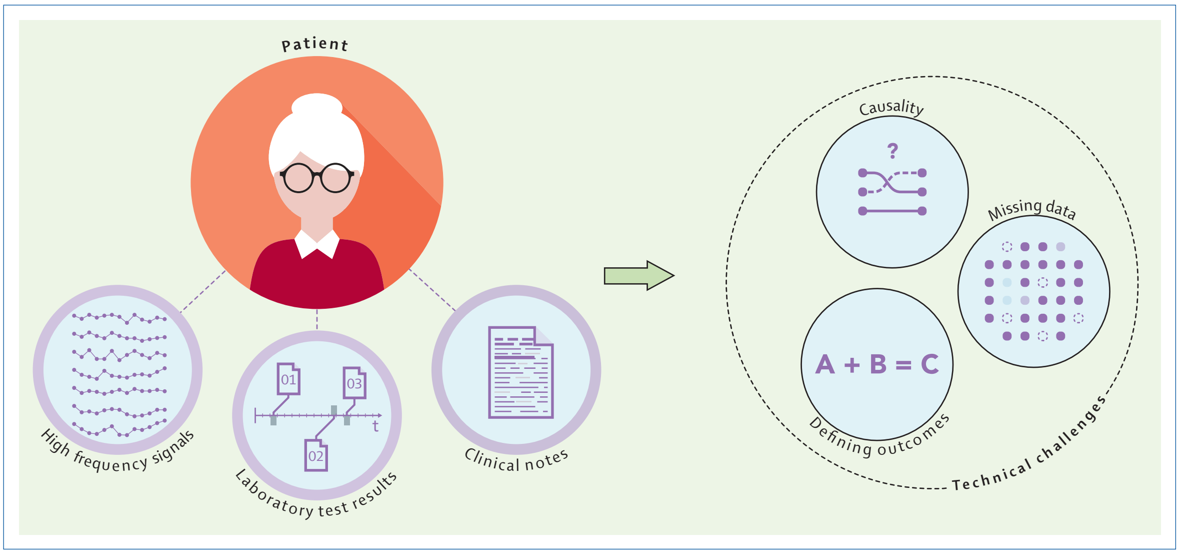

Figure: **Opportunities and challenges of using artificial intelligence in health care**
Figure reproduced with permission of Anders Häggman.

difficult in clinical notes because they often contain misspelt words, lots of acronyms, and text that has been copied and pasted. Even software packages that are designed to process clinical text can be misled by clinical notes;[4] for example, an algorithm trained on a large corpus of medical text might incorrectly identify patients with autism as having cancer, because T2 is both the clinical term for a stage of cancer progression, and an output class of MRI pulse sequences used to diagnose autism in children. Although this problem has been overcome in other settings, such as web searches, medical text is often filled with jargon and coded language.

In addition to the challenges of data heterogeneity, algorithms in health care must address well defined tasks that are clinically important, and work to identify new and important capacities. We identify three technical challenges, with plausible short-term solutions and long-term outlooks (appendix p 3).

Firstly, clinical data tend to be messy, incomplete, and potentially biased. Imputation, sparse encoding, or matrix factorisation methods (appendix p 2) can be used to address incomplete or missing data features, but they must be used with caution because the correct method depends on which data are missing. Robust inference also depends on large, representative datasets. However, gaining access to a sufficient amount of data in a health-care setting can be challenging because of patient privacy restrictions. A possible short-term solution is to use generative adversarial networks with differential privacy to generate synthetic data with the statistical properties of real health-care data; however, models trained on synthetic data might not be as accurate as those trained on clinical data.[5] A long-term solution to data challenges must focus on generating high-quality, deidentified data primarily for research purposes. Efforts such as the US National Institute of Health's All of Us Research Program and the UK Biobank have generated databases that are accessible to researchers globally. Continuing to scale up these kinds of efforts will help to alleviate many, but not all, of the issues associated with the secondary use of EHR data.

Secondly, if outcomes are predicted on the basis of measurements, problems can arise when the measurements change considerably. For example, a model trained on data from an urban hospital might not be able to predict outcomes in a rural setting. Measurements can change over time as patient populations change or care policies evolve.[6] Data might unintentionally be confounded by measurement drift (appendix) as equipment ages or changes, which can be adjusted for if the drift effect is identified. Unsupervised learning methods or causal inference approaches could potentially be used to detect shifts in the underlying population.

Ideally, models should have some ability to handle new diseases, such as the first case of Zika virus in North America. Although many state-of-the-art methods are known to be overconfident under standard training conditions, AI techniques that are forced to assign probabilities to any input can provide a foundation for understanding when a situation is unknown.[7] Handling the unknown, or at least knowing when the current situation is an unknown, is important in ensuring patient safety and clinician trust in algorithms. In the long term, regulatory incentives are needed for the creation of better devices that can expedite the acquisition and availability of clinical data. By improving the coverage of data sources, we can begin to detect conditions of interest in settings that have not previously been explored.

Finally, in the field of medicine, labels (eg, disease states) are assigned by experts, but not all experts will agree on the same label. When algorithms learn labels from data, uncertainty in our understanding of the label complicates modelling. For example, whereas some diseases (eg, diabetes) are verifiable through blood tests, others (eg, heart failure) might encompass a variety of underlying conditions, thereby requiring human judgment to label each patient.[7,8]

Label uncertainty can be addressed by using generative models and unsupervised clustering to separate populations into underlying subtypes, but this approach assumes that there are enough variables available to learn class separations. Diagnostic baselines could also be used to decide how variables should be used in models, making use of clinical knowledge, but this knowledge might progress and diagnostic criteria might change or become contentious as guidelines for medical treatments evolve.[9]

In the long term, full capture of data from robust sources is needed to match self-reported patient data with expert-verified clinical outcomes. Although previously natural divisions of expertise have occurred when working with the heterogeneous data that are generated from the health-care system, including text, speech, and images, these divisions have begun to blur as systems are increasingly using different modalities of data.

We recommend that clinicians and AI researchers work collaboratively to pair clinical challenges with novel technical solutions.[10] Engaging in close partnerships will create meaningful algorithms, foster community, and form culture.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, Rajesh Ranganath
University of Toronto and Vector Institute, Toronto, ON M5S 3G8, Canada (MG); Microsoft Research, Redmond, WA, USA (TN); Johns Hopkins University, Baltimore, MD, USA (PS); Harvard School of Public Health, Boston, MA, USA (ALB); Massachusetts Institute of Technology, Cambridge, MA, USA (IYC); and New York University, New York, NY, USA (RR)
marzyeh@cs.toronto.edu

PS reports personal fees and other fees from Bayesian Health, outside the submitted work. All other authors declare no competing interests.

1 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; **25:** 44–56.

2 Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. *JAMA* 2014; **311:** 2479–80.

3 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; **361:** k1479.

4 Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; **17:** 507–13.

5 Beaulieu-Jones BK, Wu ZS, Williams C, et al. Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv* 2018; published online Dec 20. DOI:10·1101/159756 (preprint).

6 Eneanya ND, Yang W, Reese PP. Reconsidering the consequences of using race to estimate kidney function. *JAMA* 2019; published online June 6. DOI:10·1001/jama.2019·5774.

7 Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. 2016 IEEE Symposium on Security and Privacy; San Jose, CA; 2016. DOI:10.1109/SP.2016.41.

8 Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. *Nat Rev Cardiol* 2011; **8:** 30–41.

9 Boyce PM, Koloski NA, Talley NJ. Irritable bowel syndrome according to varying diagnostic criteria: are the new Rome II criteria unnecessarily restrictive for research and practice? *Am J Gastroenterol* 2000; **95:** 3176–83.

10 Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, et al. Making big data useful for health care: a summary of the inaugural MIT critical data conference. *JMIR Med Inform* 2014; **2:** e22.