The myth of generalisability in clinical research and machine learning in health care

Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, Leo Anthony Celi**

*Joint senior authorship

School of Engineering & Applied Sciences, Harvard University, Cambridge, MA, USA (J Futoma PhD, F Doshi-Velez PhD); Department of Medicine, NYU Langone Health, New York, NY, USA (M Simons MD); Department of Health Policy and Management (T Panch MD) and Department of Biostatistics (L A Celi MD), Harvard T.H. Chan School of Public Health, Boston, MA, USA; Wellframe, Boston, MA, USA (T Panch); Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA (L A Celi); and Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA (L A Celi)

Correspondence to:
Dr Leo Anthony Celi, Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA
lceli@bidmc.harvard.edu

Lancet Digital Health 2020; 2: e489–92

An emphasis on overly broad notions of generalisability as it pertains to applications of machine learning in health care can overlook situations in which machine learning might provide clinical utility. We believe that this narrow focus on generalisability should be replaced with wider considerations for the ultimate goal of building machine learning systems that are useful at the bedside.

Introduction

Dr Lee, an esteemed intensivist from the USA, is rounding in an intensive care unit (ICU). He is asked by a team member who is taking care of patients with COVID-19 if they can triage their patients to optimise use of scarce resources, such as ventilators, with their hospital's new machine learning model to predict mortality.[1] He is about to say yes, but stops himself. Do the findings of the preprints and fast-tracked published articles that this model is based on apply to his patient population?[2] Problems with the increase in hastily written articles notwithstanding, are the conclusions of research based on patients with COVID-19 in China and Italy from several months ago still valid in his ICU today, given the differences in practice patterns and rapidly changing guidelines and protocols?

The answers to these questions strongly depend on context. For a substantial number of individuals who die in the ICU, their mortality is a result of cessation of treatment. Many factors affect the decision to discontinue invasive interventions, including whether the outcome is aligned with the patient's preferences. Therefore, a machine learning model that predicts hospital mortality is largely identifying which patients are most likely to discontinue treatment, and is effectively learning a collection of rules to predict this outcome.

As a thoughtful clinician, Dr Lee realised that he should consider the broader context in which the ICU's mortality model was developed. Unlike Dr Lee, current machine learning systems typically cannot identify differences in contexts, let alone adapt to them.[3] However, the collections of rules derived by machine learning systems might still be effective in a specific context. In this Viewpoint, we argue that an overemphasis on overly broad notions of generalisability overlooks situations in which machine learning systems have the greatest ability to deliver clinical utility.

In pursuit of generalisability

Generalisability is not a binary concept, and does not have a universally agreed definition. According to one common hierarchy,[4] a set of rules from a machine learning system or a clinician might be applicable: internally, applying only in the narrow context in which it was developed; temporally, applying prospectively at the centre in which it was developed; or externally, applying both at new centres and in new time periods. Other hierarchies construct even more detailed levels of generalisability.[5]

A system that achieves the highest possible level of generalisability is desirable. Many medical journals mandate that articles on machine learning applications show results on external cohorts.[6–8] This request is only natural: such journals often have diverse readerships, and research articles with widespread relevance to many readers are more likely to be read and circulated, increasing the visibility of the journal and publishers. Similarly, vendors, such as electronic health record companies, prize generalisability in their applications. These companies frequently sell generic black-box machine learning systems that purport to apply universally across many hospitals.[9] Broad applicability is to their financial advantage as it allows for amortising development costs for machine learning and hopefully eliminates the need for solutions tailored to each hospital.

In some cases, such broad geographical generalisability might be feasible—eg, in medical imaging applications such as diagnosing diabetic retinopathy.[10] However, these areas are still not immune to generalisability issues,[11] and few prospective studies or randomised trials exist.[12] More often, universality is a myth. As users of these machine learning systems can attest, the demand for universal rules—generalisability—often results in systems that sacrifice strong performance at a single site for systems with mediocre or poor performance at many sites.[13–15] The inherent trade-off that clinicians and researchers alike encounter is between improving system performance locally and having systems that generalise.

Although we have already explored the story of Dr Lee, let us discuss a more general hypothetical scenario. Consider a machine learning system built by tertiary care hospital A to help clinicians identify patients who are at high risk of a hospital-acquired, highly contagious diarrhoeal infection. A prospective study at hospital A found the system was effective at helping infection-control practitioners prevent outbreaks, and the system was put into general use. Hearing of this success, their partner rural-community hospital, hospital B, decided to adopt the system. Unfortunately, its performance at hospital B was poor. Investigating the problem revealed that there was an antibiotic stewardship in hospital A, but not in hospital B. The machine learning system, implicitly trained for a context in which a certain policy around antibiotics exists, was unusable at hospital B. Does this mean it should not be used at hospital A?

Of course not: the machine learning system had already shown temporal generalisability at hospital A, providing tailored predictions that ensured cases did not go unnoticed. Geographical generalisation to hospital B is not necessary for clinicians at hospital A to use the system to improve patient care. Rather, the desire for geographical generalisability is a proxy for validity: a theoretical machine learning system applied universally would tautologically always work as expected. In situations with strong signals and little local variability across sites, this strategy might make sense. However, in many scenarios there are too many practice patterns and other local idiosyncrasies that make learning a broadly applicable model effectively impossible. Instead, machine learning systems in these settings can be viewed as an aspirational form of evidence-based medicine—local data to create local inferences for local patients and clinicians.[16–19] Issues of generalisability are not unique to machine learning and are a dominant concern for clinical guidelines where the results of randomised controlled trials, the gold standard for evidence generation, might not generalise beyond the trial settings.[20–23] If hospitals want to have useful machine learning systems at the bedside, the broader research community need to stop focusing solely on generalisability and consider the ultimate goal: will this system be useful in this specific case?

Panel: Overview of potential threats to generalisability in clinical research and machine learning in health care, along with hypothetical examples of what they might look like in practice

Changes in practice pattern over time

- Improved patient outcomes through adoption of low-tidal-volume ventilation in the intensive care unit (ICU) will affect the performance of models that were developed when higher tidal volumes were standard.
- Leucodepletion of blood for transfusion became standard of care in most countries. Models related to blood transfusion and outcomes require recalibration if validated before the practice change.

Differences in practice between health systems

- Mortality predictions for patients admitted to the ICU with COVID-19 are highly sensitive to criteria for ICU admission across hospitals, which in turn vary depending on ICU demand and capacity.

Patient demographic variation

- Models to predict the risk of hospitalisation from COVID-19 that are trained on data from Italy where there is a high proportion of older individuals in the population will not do well in countries with a different age distribution—eg, low-income and middle-income countries that typically have a younger population.

Patient genotypic and phenotypic variation

- Model performance is linked to the composition of the training cohort with regard to disease genotypes or phenotypes, or both. These models will not translate well to populations in which the genotypic or phenotypic make-up is different. Some phenotypes of sepsis and acute respiratory distress syndrome, for example, might be over-represented or under-represented in different settings.

Hardware and software variation for data capture

- Bedside monitors that have different sampling rates for the capture of physiological signals and that are measured continuously will have different susceptibilities to artifacts and will affect models that have time-series data as an input.
- Computer-vision models for automated interpretation of CT scans are sensitive to the machines used to obtain the images.

Variation in other determinants of health and disease (eg, environmental, social, political, and cultural)

- A model developed in the USA to predict neurological outcomes of premature babies will not do well in a low-income country because of resource availability.
- The relationship of patient and disease factors with clinical events, such as hospital-acquired infection, will change when a health-care system is strained (eg, during a pandemic).

Beyond generalisability

To create machine learning systems that are clinically useful, the emphasis should shift from demanding geographical generalisability to understanding how, when, and why a machine learning system works. This knowledge will help medical professionals use the system correctly, not only across institutions but also within an institution as patients and practices change. For instance, if hospital A stopped their antibiotic stewardship policy, they should know to update their machine learning system. Although the precise level of generalisability required for a real-world application will depend on the context, any system intended to be integrated into a clinical environment will need to be at least temporally generalisable, ensuring that it performs well prospectively. We further suggest that there are several important questions to ask when assessing the overall validity of a machine learning system for a particular context: when the machine learning system is right, is it right for the right reasons? Or is it relying on anticausal mechanisms due to unobserved confounders (as in the example of patients with asthmatic pneumonia who have lower mortality rates than people who do not have asthma because of more intensive care[24])? How do the characteristics of the cohort used to develop the machine learning system compare with typical patients at the institution where it will be used? Does the system rely on variables known to be collected differently at different centres?

More broadly, all machine learning systems must be closely monitored to make sure that their performance does not degrade with time as patient demographics and practice patterns inevitably shift.[25,26] Furthermore, techniques from continual learning[27] offer enormous potential to create more advanced machine learning systems that continuously update based on new data. In theory, such systems could address many of the pitfalls with generalisability (panel). However, these types of self-updating algorithms pose enormous regulatory challenges, as outlined in a recent white paper by the US Food and Drug Administration,[28] and there are still many technical and cultural barriers to integrate them into real-world systems.[29]

This path to validation will probably require more work than simply evaluating a machine learning system on multiple datasets and then claiming universal external generalisability, as vendors might wish to do. However, models thoroughly vetted through these proposed standards have the promise of being able to do better for the context at hand. We also emphasise that the methodological process of developing a high-quality machine learning system might be generalisable: the lessons hospital A learns about how to prepare data and then train, test, and monitor their machine learning system can be used by hospital B to do the same with their own data. As we gain a deeper understanding of what patterns different machine learning systems rely on, we can also determine more accurately to what extent systems trained in one context might work in another—eg, perhaps hospital A's machine learning system would also work well at hospital C, a neighbouring tertiary care institution that follows very similar practices. Where possible, multicentre datasets might offer the potential to better capture heterogeneity across sites during model development, potentially leading to more generalisable models. Multicentre data from the relevant target populations are also the only way to validate whether a model truly generalises to a new institution.[30,31]

Finally, we note that there are some circumstances where broad generalisability is desirable. For example, if we are interested in using machine learning systems to understand the underpinnings of disease (eg, a study to identify biomarkers that predict which patients with COVID-19 will develop cytokine storm), then the machine learning system's output should not be influenced by practice-specific variables, such as the specific technology used to take measurements. However, clinicians do constantly adjust their behaviours and practices depending on the unique characteristics of patients, the availability of resources, and the local practice norms. If we are to build accurate and actionable machine learning systems, we should not ignore the fact that practice-specific information is often highly predictive.[32–34]

Conclusion

Machine learning systems are not like thermometers, reliably measuring the temperature via universal rules of physics; nor are they like trained clinicians, gracefully adapting to new circumstances. Rather, these systems should be viewed as a set of rules that were trained to operate under certain contexts and rely on certain assumptions, and might work seamlessly at one centre but fail altogether somewhere else. We hope this Viewpoint will help reframe the narrow focus on generalisability and will encourage future researchers, developers, and reviewers to be explicit about the appropriate level of generalisability for their setting. We believe that a renewed focus on broader questions about characterising when, how, and why machine learning systems have clinical utility will help ensure that these systems work as intended for both clinicians and for patients.

Contributors

All authors contributed to the writing of the initial manuscript and to the editing process.

Declaration of interests

TP reports stock and is a paid employee of Wellframe, outside of the submitted work. FD-V reports grants from the National Institutes of Health, and personal fees from Davita Kidney Care and Google Health via Adecco, outside of the submitted work. JF reports grants from the National Science Foundation, outside of the submitted work. LAC reports grants from the National Institutes of Health, outside of the submitted work. MS declares no competing interests.

References

1 Truog RD, Mitchell C, Daley GQ. The toughest triage—allocating ventilators in a pandemic. *N Engl J Med* 2020; **382:** 1973–75.

2 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.

3 Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: a survey on few-shot learning. *arXiv* 2020; published online March 29. https://doi.org/10.1145/3386252 (preprint).

4 Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; **19:** 453–73.

5 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; **130:** 515–24.

6 Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology* 2020; **294:** 487–89.

7 Leisman DE, Harhay MO, Lederer DJ, et al. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020; **48:** 623–33.

8 Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016; **18:** e323.

9 Ross C. Hospitals are using AI to predict the decline of Covid-19 patients—before knowing it works. April 24, 2020. Stat. https://www.statnews.com/2020/04/24/coronavirus-hospitals-use-ai-to-predict-patient-decline-before-knowing-it-works (accessed May 31, 2020).

10 Gulshan V, Rajan RP, Widner K, et al. Performance of a deep-learning algorithm *vs* manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol* 2019; **137:** 987–93.

11 Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; **15:** e1002683.

12 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **368:** m689.

13 Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Crit Care Med* 2019; **47:** 49–55.

14 Downey CL, Tahir W, Randell R, Brown JM, Jayne DG. Strengths and limitations of early warning scores: a systematic review and narrative synthesis. *Int J Nurs Stud* 2017; **76:** 106–19.

15 Gerry S, Bonnici T, Birks J, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020; **369:** m1501.

16 Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018; **24:** 1716–20.

17 Hampton JR. Evidence-based medicine, opinion-based medicine, and real-world medicine. *Perspect Biol Med* 2002; **45:** 549–68.

18 Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence—what is it and what can it tell us? *N Engl J Med* 2016; **375:** 2293–97.

19 Panch T, Pollard TJ, Mattie H, Lindemer E, Keane PA, Celi LA. "Yes, but will it work for *my* patients?" Driving clinically relevant research with benchmark datasets. *NPJ Digit Med* 2020; **3:** 87.

20 Gluud LL. Bias in clinical intervention research. *Am J Epidemiol* 2006; **163:** 493–501.

21 Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001; **323:** 42–46.

22 Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials* 2006; **1:** e9.

23 Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *Lancet* 2005; **365:** 82–93.

24 Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997; **9:** 107–38.

25 Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017; **24:** 1052–61.

26 Nestor B, McDermott MBA, Boag W, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *Proc Mach Learn Res* 2019; **106:** 1–23.

27 Parisi GI, Kemker R, Part JL, Kanan C, Wermter S. Continual lifelong learning with neural networks: a review. *Neural Netw* 2019; **113:** 54–71.

28 US FDA. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. 2019. https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf (accessed May 31, 2020).

29 Lee CS, Lee AY. Clinical applications of continual learning machine learning. *Lancet Digit Health* 2020; **2:** e279–81.

30 Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. *Diagn Progn Res* 2019; **3:** 6.

31 Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353:** i3140.

32 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.

33 Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *J Am Med Inform Assoc* 2014; **21:** 699–706.

34 Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: observational study. *JMIR Med Inform* 2019; **7:** e11605.