nature medicine

PERSPECTIVE

https://doi.org/10.1038/s41591-019-0548-6

Do no harm: a roadmap for responsible machine learning for health care

Jenna Wiens[1,19*], Suchi Saria[2,3,4,19], Mark Sendak[5], Marzyeh Ghassemi[6,7,8], Vincent X. Liu[9], Finale Doshi-Velez[10], Kenneth Jung[11], Katherine Heller[12,13], David Kale[14], Mohammed Saeed[15], Pilar N. Ossorio[16], Sonoo Thadaney-Israni[17] and Anna Goldenberg[6,8,18,19*]

Interest in machine-learning applications within medicine has been growing, but few studies have progressed to deployment in patient care. We present a framework, context and ultimately guidelines for accelerating the translation of machine-learning-based interventions in health care. To be successful, translation will require a team of engaged stakeholders and a systematic process from beginning (problem formulation) to end (widespread deployment).

The potential impact of machine learning (ML) in health care warrants genuine enthusiasm, but its limited adoption in clinical care to date indicates that many of the current strategies are far from optimal. Although successful translation requires bringing together expertise and stakeholders from many disciplines, the development of ML solutions is currently occurring in silos. Past work has tackled particular aspects of data-analysis challenges: separating causation from correlation[1], identifying lurking biases in data[2] and regulating predictive analytics[3]. Here, we take a step back, providing a comprehensive overview of the barriers to deployment and translational impact. With a view toward accelerating safe, ethically responsible and meaningful progress in ML for health care, we lay out critical steps to consider when designing, testing and deploying new solutions. ML deployment in any field should be carried out by interdisciplinary teams including knowledge experts, decision-makers and users (Table 1). Accordingly, this roadmap is intended for a broad audience, while making specific recommendations for critical contributors to such initiatives.

Choosing the right problems

Progress in ML for health care to date has been limited by the lack of well-defined questions and a dearth of annotated datasets. Many ML researchers remain focused on questions for which annotations are readily available, without necessarily questioning the clinical relevance of the problems and their solutions. For example, a popular benchmark challenge in the community focuses on predicting in-hospital mortality on the basis of data collected during the first 48 hours after admission to the intensive care unit[4]. Clearly annotated data are publicly available, and in recent years, performance on this task has approached an area under the curve (Box 1) of 0.9 (ref. [5]). However, assessing clinical utility requires careful evaluation against the scenario in which the model will be used. For example, a model may learn to associate patterns of end-of-life care with a high risk of mortality; as a result, despite the high area under the curve,

Box 1 | Glossary of ML terms

Area under the curve: short for 'area under the receiver operating characteristics curve', a measure of discriminative performance that summarizes the trade-off between sensitivity and specificity

Generalize: the ability to generate useful estimates in new never-before-seen examples (for example, patients from another hospital)

Ground truth: information provided by direct observation, often used to refer to the training labels in the context of ML

Label leakage: when the labels (that is, outcomes) of interest are erroneously (perhaps implicitly) included in the input

Model: a learned mapping from some input (for example, covariates representing a patient) to some output (for example, risk of mortality)

Operating regime: most models output continuous estimates that can then be thresholded; the choice of threshold corresponds to a specific sensitivity, specificity, positive predictive value and so forth, and reflects the operating regime

Stepped-wedge trial: participants receive treatments in 'waves' rather than complete randomization (for example, intervention is applied gradually one unit a time)

Training: the process of learning the mapping that constitutes the model

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA. [2]Departments of Computer Science and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Health Policy and Management, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [4]Bayesian Health, New York, NY, USA. [5]Duke Institute for Health Innovation, Duke University School of Medicine, Durham, NC, USA. [6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [7]Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [8]Vector Institute, Toronto, Ontario, Canada. [9]Kaiser Permanente Division of Research, Oakland, CA, USA. [10]School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA. [11]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA. [12]Google Inc., Mountain View, CA, USA. [13]Department of Statistical Science, Duke University, Durham, NC, USA. [14]Information Sciences Institute, University of Southern California, Los Angeles, CA, USA. [15]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [16]Law School, University of Wisconsin–Madison, Madison, WI, USA. [17]Presence and Program in Bedside Medicine, Stanford University School of Medicine, Stanford, CA, USA. [18]SickKids Research Institute, Toronto, Ontario, Canada. [19]These authors contributed equally: Jenna Wiens, Suchi Saria, Anna Goldenberg. *e-mail: wiensj@umich.edu; anna.goldenberg@utoronto.ca

Table 1 | Interdisciplinary teams may consist of stakeholders from different categories

Stakeholder categories	Examples
Knowledge experts	• Clinical experts • ML researchers • Health information and technology experts • Implementation experts
Decision-makers	• Hospital administrators • Institutional leadership • Regulatory agencies • State and federal government
Users	• Nurses • Physicians • Laboratory technicians • Patients • Friends and family (framily)

such an alert would not be clinically useful because the model is not conveying anything that the care team does not already know. If data on whether a patient has been clinically determined to be near end of life were explicitly charted, then such information could explicitly be taken into account to improve the clinical utility of the derived model. However, a lack of intimate knowledge of the data often leads to misguided problem formulations and solutions. When starting new projects, such issues can be avoided by engaging relevant stakeholders before developing solutions. For ML researchers, engaging stakeholders early in any project can help identify the right problem to focus on. Beyond clinicians, key stakeholders may include financial and operational administrative leaders, patients, friends and family ('framily'), and insurers, when relevant. It is important to remember that the most impactful problems are those that have not only clinical relevance but also champions throughout the various stages of the development and deployment framework.

Developing a useful solution

After the problem is formulated, and permissions regarding data use and access are in place, one may be eager to start developing a solution. But before a model is trained, the data must be thoroughly scrutinized to ensure that they are appropriate for the problem in question. First, when predicting a specific outcome, one must ask the questions of how and when the data were collected and what purpose the data serve. These questions apply to both the outcome (that is, model output) and predictors (that is, model input). For example, international classification of diseases (ICD) codes, entered into the electronic health record after a patient's visit, are primarily used for billing purposes. They may poorly correspond to the actual diagnoses[6] and should not be relied on as ground truth. However, they could be used as input to the model if they are available at prediction time. Second, one must understand and account for differences in the ways in which data are collected. For a model to operate across different departments or health systems, the underlying data elements should be harmonized. Here, custom piecemeal solutions developed by individual researchers could benefit from top-down policies that facilitate harmonization across systems. Harmonization can ensure that data-collection formats or frequency do not vary in unanticipated ways. Third, the training data must critically be representative of the environment where the model will be used. Data for certain populations may be underrepresented. ML researchers must ensure that adequate data are present if the model is intended for use in these populations. Finally, there are many ways in which subtle biases in electronic health data can decrease model reliability, especially if these biases are not corrected for during model development[7,8]. For example, models may identify department- or provider-specific patterns as predictors. These predictors will not generalize as practice patterns change or as the model is applied across care environments[9]. Measuring and accounting for these biases requires deeper understanding of changes in the data that pose a threat to model correctness; such issues should be identified up front and ideally corrected during model development[7].

Considering ethical implications

Numerous researchers have identified ways in which non-health-related ML can exacerbate existing social inequalities by reflecting and amplifying existing race, sex and other biases[10]. Health care is not immune to pernicious bias[11]. The health data on which algorithms are trained are likely to be influenced by many facets of social inequality, including bias toward those who contribute the most data. For example, algorithms for predicting whether an individual should receive a surgery may be biased toward those who are able to access and afford the procedure. Some of this bias can be corrected for during model training[7], but, in general, awareness is necessary to investigate when potential biases are lurking in the data and what can be done to mitigate their effect. Beyond model performance, there may be ethical questions raised by the problem statement itself. For example, smoking status is a useful predictor for many outcomes (for example, lung cancer). Using data-mining techniques, researchers could often infer or impute smoking status if missing, but is it ethical to do so for subjects who declined to provide this information? What about imputing HIV status? Ethicists, social scientists and regulatory scholars working with ML experts and other relevant stakeholders can help a project team think about bias and how to counter it well before it is incorporated into the model. Here, the benefits go both ways: ethicists can also benefit from engaging with ML experts to fully grasp how ML techniques can either amplify or combat biases. Such partnerships can ensure that the project respects participants' rights and interests. If the ultimate goal is to deploy an ML tool in clinical practice, researchers must consider ethical implications up front to promote the privacy, safety and fair treatment of patients and all parties affected by the tool.

Rigorously evaluating the model

Beyond ethical, legal and moral challenges, there are technical issues to consider when evaluating a model's readiness for use. When a model is tuned to a given environment, validation within that environment requires careful thought to ensure that no unintended label leakage has occurred between the datasets used for model tuning and independent testing. Consider, for instance, the 'radiologist-level' performance recently achieved across several tasks using chest X-rays[12]. The data used in the analysis consisted of multiple frontal-view X-ray images per patient. It was important to split data at the patient level, as opposed to random splitting, so that no images from the same patient appeared in both the training and testing sets. Label leakage may arise in even more subtle or unexpected ways. For example, in a diagnostic task using speech, if significant class imbalance exists across sites, then a model may learn to classify data-collection sites (for example, on the basis of background noise) rather than to diagnose. Second, evaluating and reporting the scope in which the model is likely to succeed and fail is important; for example, sepsis-prediction models developed on adults are likely not to be applicable to a pediatric population, owing to differences in pediatric sepsis presentation. Third, statistical analysis should consider clinically relevant evaluation metrics, as recently described[3]. The commonly used area under the curve integrates over all error regimes and consequently may be irrelevant for much of clinical practice. For example, in predicting a relatively rare outcome (such as a healthcare-associated infection with *Clostridium difficile*), operating in a regime that corresponds to a high false-positive rate may be impractical, because costly interventions might be applied in situations in which patients are unlikely to benefit. Measures such as positive predictive value and sensitivity at reasonable

operating points are often more informative. In many cases, measures of cost of deployment or impact (for example, on clinical and financial outcomes) might also be considered. If the effectiveness and cost of the intervention are known, the 'number needed to benefit' can also be calculated as the number of patients that the model must flag to identify one true-positive patient, multiplied by the number of true-positive patients who must be treated for one patient to benefit[13].

Beyond quantitative measures of performance, qualitative approaches can expose concerns associated with bias and confounding that the quantitative measures might have missed. For example, clinical experts can investigate explanations provided at individual test points to determine whether the model is plausible and relevant. In one study, through reviewing a list of medications associated with a high risk of developing *C. difficile* infections, clinicians identified oral vancomycin. Oral vancomycin is a treatment for infection, thus suggesting that the model had learned an association between empiric treatment and outcome. After this issue was identified, it was easily addressed by including only data up to 2 days before a diagnosis during model training[14]. A lack of explainability does not mean that a model is useless. However, a clearly flawed explanation (such as empiric treatment, as in the example above) may mean that the model in its present form is not adequate for the task for which it was intended and should be retrained to avoid the identified issue before use. Further, methods for obtaining measures for how reliable a model's predictions are at a given test point can be used to guide qualitative evaluation in regions of greater uncertainty[15].

Thoughtfully reporting results

ML researchers must be mindful when interpreting and reporting their results. Proposed reporting guidelines developed by the community provide a good starting point for the information to include in a manuscript[11]. Such guidelines outline the importance of clear descriptions of the source of the data, participants, outcomes and predictors, and in some cases require the model itself (that is, all regression coefficients) to be presented. This last requirement creates a potential for unintended consequences and even harm, if the model is then applied inappropriately. For example, a recent study in building models to predict healthcare-associated infections found that variables associated with risk at one hospital were protective in another[14]. Thus, it is imperative that authors report the context(s) in which the model applies and was validated, including a discussion of indications and contraindications for deployment, and what assumptions or conditions should be satisfied. Second, it is good practice to share code, packages and inputs used to generate the reported results, as well as supporting documentation. Recent work has demonstrated that deep neural networks (and other high-capacity models) can generate vastly different results solely on the basis of the random seed used to initialize the optimization procedure[16]. Sharing code can help the community better understand and build on past work. Finally, when comparing multiple models, researchers must go beyond predictive performance. For example, it would be informative to include an analysis of the trade-offs between simpler, faster and more explainable models versus complex, slower but more accurate models. Following these and existing reporting guidelines would help facilitate reproducibility, accelerate science and limit the possibility of misinterpreted conclusions and misapplied tools.

Deploying responsibly

Effectively applying a predictive model in an ethical, legal and morally responsible manner within a real-world healthcare setting can be substantially more difficult than developing a model in a curated experimental environment. Before integrating in patient care, it is critical to test the system in 'silent' mode, in which predictions are made in real time and exposed to a group of clinical experts but not acted upon. This prospective validation allows clinicians to identify and review errors in real time[17]. After prospective validation, the efficacy of the model in clinical studies can be evaluated. Randomized controlled trials are often the cleanest way of evaluating efficacy. However, patient-level or physician-level randomization can be challenging in many settings. For example, in deploying an early-warning system for sepsis, an ML-based system requires a novel workflow, such as presenting a display that indicates why the model generated an alert, to facilitate evaluation[18]. Randomization across patients would require clinicians to use different workflows depending on whether the patient is assigned to the control or the treatment arm. This process can confuse providers and might be perceived by hospital administrators as a source of risk to patient safety. As a result, randomization across patients for ML interventions that require changes to workflow is not feasible. Pre–post studies with seasonality corrections and stepped-wedge trials[19] that introduce changes gradually across an institution may be good potential alternatives. In addition to trial design, understanding how to best integrate the intervention with the care team's workflow is equally important: how and to whom should the output of the model be presented? Even the best-performing models may fail to have an impact if these questions are not properly addressed[3]. By investing in infrastructure for prospective evaluation and implementation of ML models, health-system leaders can accelerate researcher-driven solutions by facilitating effective model deployment. Finally, populations and clinical protocols change over time, and frequent monitoring to assess for pointwise reliability and errors can aid in discovering opportunities for improvement[15]. Because ML systems can learn from data to improve performance over time, the US Food and Drug Administration has recently begun investigating the establishment of guidelines to allow for continuous improvement while ensuring efficacy and safety. This is an active area of research and policy regulation. As systems are deployed across an institution, centralized resources for maintaining model safety are necessary to ensure the wide operational success of the project.

Making it to market

To transition to the market, ML tools must be validated with the government-required regulatory steps in mind, specific to the country of deployment. For example, in the United States, some types of medical software or clinical decision support systems (CDS) are considered and regulated as medical devices, but others are not[20]. Four criteria determine whether CDS software are regulated as medical devices, the two most important of which are first that the software cannot receive, analyze or otherwise process a medical image or signal from an in vitro diagnostic device (such as a DNA-sequencing machine) or from any other signal acquisition system, and second that a healthcare professional must be able to understand the basis of its recommendations. The software also cannot be intended as the sole source of recommendations regarding treatment, diagnosis or prevention of a disease[21]. Consequently, developers and designers must be very careful about what types of information they provide back to clinical decision-makers. Developers must also consider whether and how the model might be updated over time. As mentioned above, the US Food and Drug Administration has recently proposed a regulatory framework for modifications to artificial intelligence/ML-based software as medical devices[22]. Other regulatory aspects of the tool may relate to the ability of the user to interrogate the model. Indeed, models that can explain their predictions may be preferred. If investors and any firms that might market an ML-based product for health care are not clear about the regulatory pathway to market, the product might not ever benefit patients. Thus, teams developing ML tools for health care must seek regulatory advice early in the process.

Conclusions

Many complexities exist in developing and deploying effective ML systems across domains as diverse as health care, self-driving cars[23]

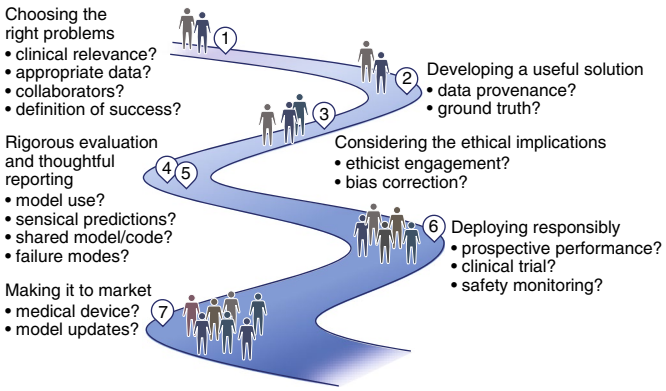

Fig. 1 | A roadmap for deploying effective ML systems in health care. By following these steps and engaging relevant stakeholders early in the process, many issues stemming from the complexity of adopting ML in practice can be successfully avoided.

and space exploration[24]. We propose a robust overarching framework (Fig. 1), including people and processes, within which many issues stemming from the complexity of adopting ML in practice, especially in health care, can be successfully avoided. We encourage all researchers working on deploying ML systems to consider the broader context early in the process.

There are many unanswered questions remaining in the field of ML: how much accuracy is sufficient for deployment? What level of model transparency is required? Do we understand when the model outputs are likely to be unreliable and therefore should not be trusted? Although there is still a long way to go, impactful innovation with ML in health care is now clearly a team activity. Ultimately, meaningful progress will require both system-wide changes driven by policy-makers and health-system leaders, and individual researchers working on bottom-up solutions. From understanding the data to deploying the model, to have the greatest impact, developers must work as well-informed, interdisciplinary teams including health-system leaders, frontline providers and patients. As the field hurtles forward, it is worth pausing to remember the Hippocratic oath: "first, do no harm." It is imperative that all stakeholders work together to understand and appropriately address the nuances and potential biases lurking in health data, before deploying solutions. Doing so will not only decrease the potential for unintended consequences but also reduce rather than amplify existing social inequalities and ultimately lead to better care.

Received: 10 July 2019; Accepted: 17 July 2019;
Published online: 19 August 2019

References

1. Lazer, D., Kennedy, R., King, G. & Vespignani, A. Big data. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
2. Hutson, M. Even artificial intelligence can acquire biases against race and gender. *Science* https://doi.org/10.1126/science.aal1053 (2017).
3. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
4. Silva, I., Moody, G., Scott, D. J., Celi, L. A. & Mark, R. G. Predicting in-hospital mortality of ICU patients: the Physionet/Computing in Cardiology Challenge 2012. *Comput. Cardiol.* **39**, 245–248 (2012).
5. Luo, Y., Cai, X., Zhang, Y. & Xu, J. Multivariate time series imputation with generative adversarial networks. in *Advances in Neural Information Processing Systems* 1596–1607 (NeurIPS, 2018).
6. O'Malley, K. J. et al. Measuring diagnoses: ICD code accuracy. *Health Serv. Res.* **40**, 1620–1639 (2005).
7. Saria, S. & Subbaswamy, A. Tutorial: safe and reliable machine learning. Preprint at https://arxiv.org/abs/1904.07204 (2019).
8. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care? *AMA J. Ethics* **21**, E167–E179 (2019).
9. Schulam, P. & Saria, S. Reliable decision support using counterfactual models. in *Advances in Neural Information Processing Systems* 1697–1708 (NeurIPS, 2017).
10. O'neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016).
11. Williams, D. R., Mohammed, S. A., Leavell, J. & Collins, C. Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Ann. NY Acad. Sci.* **1186**, 69–101 (2010).
12. Rajpurkar, P. et al. Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at https://arxiv.org/abs/1711.05225 (2017).
13. Liu, V.X., Bates, D.W., Wiens, J. & Shah, N.H. The number needed to benefit: estimating the value of predictive analytics in healthcare. *J. Am. Med. Inform. Assoc.* https://doi.org/10.1093/jamia/ocz088 (2019).
14. Oh, J. et al. A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect. Control Hosp. Epidemiol.* **39**, 425–433 (2018).
15. Schulam, P. & Saria, S. Can you trust this prediction? Auditing pointwise reliability after learning. in *The 22nd International Conference on Artificial Intelligence and Statistics* 1022–1031 (PMLR, 2019).
16. Henderson, P. et al. Deep reinforcement learning that matters. in *Thirty-second AAAI Conference on Artificial Intelligence* (AAAI, 2018).
17. Nestor, B. et al. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. Preprint at https://arxiv.org/abs/1811.12583 (2018).
18. Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* **7**, 299ra122 (2015).
19. Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J. & Lilford, R. J. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *Br. Med. J.* **350**, h391 (2015).
20. Evans, B. & Ossorio, P. The challenge of regulating clinical decision support software after 21st century cures. *Am. J. Law Med.* **44**, 237–251 (2018).
21. Okoro, A. O. Preface: The 21st Century Cures Act—a cure for the 21st century? *Am. J. Law Med.* **44**, 155 (2018).
22. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)* (U.S. Food & Drug Administration, 2019); https://www.fda.gov/media/122535/download
23. Massachusetts Institute of Technology. Self-driving cars, robots: identifying AI 'blind spots'. *ScienceDaily* (25 January 2019).
24. Chien, S. & Wagstaff, K. L. Robotic space exploration agents. *Sci. Robot.* **2**, eaan4831 (2017).

Acknowledgements

The authors would like to thank the participants in the MLHC Conference 2018 (http://www.mlforhc.org), specifically the organizers and participants of the pre-meeting workshop that served as the genesis for this manuscript, for providing valuable feedback on the initial ideas through a panel discussion.

Competing interests

J.W., F.D.-V., D.K. and K.J. are on the board of Machine Learning for Healthcare, a non-profit organization that hosts a yearly academic meeting; they are reimbursed for registration and travel expenses. F.D.-V. consults for DaVita, a healthcare company. S.T.-I. serves on the board of Scients (https://scients.org/) and is reimbursed for travel expenses. S.S. is a founder of, and holds equity in, Bayesian Health. The results of the study discussed in this publication could affect the value of Bayesian Health. This arrangement has been reviewed and approved by Johns Hopkins University in accordance with its conflict-of-interest policies. S.S. is a member of the scientific advisory board for PatientPing. M. Sendak is a named inventor of the Sepsis Watch deep-learning model, which was licensed from Duke University by Cohere Med, Inc. M. Sendak does not hold any equity in Cohere Med, Inc. M. Saeed is a founder and Chief Medical Officer at HEALTH at SCALE Technologies and holds equity in this company. P.O. consults for Roche-Genentech, from whom she has received travel reimbursement and consulting fees of less than $4,000/year. A.G., K.H., M.G. and V.L. have no conflicts to declare.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to J.W. or A.G.

Peer review information: Joao Monteiro was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.