Patterns

CellPress
OPEN ACCESS

Preview

Can an accurate model be bad?

Melissa D. McCradden,[1,2,3,*] Mjaye L. Mazwi,[4,5] and Lauren Oakden-Rayner[1]
[1]Australian Institute for Machine Learning, Adelaide, SA, Australia
[2]Women's and Children's Health Network, Adelaide, SA, Australia
[3]SickKids Research Institute, Toronto, ON, Canada
[4]Seattle Children's Hospital, Seattle, WA, USA
[5]Department of Pediatrics, University of Washington, Seattle, WA, USA
*Correspondence: melissa.mccradden@adelaide.edu.au
https://doi.org/10.1016/j.patter.2025.101205

Outcome-prediction models can harm patients even when they have good accuracy, as shown in a recent *Patterns* paper by Van Amsterdam et al. In this preview, we consider the ethical and empirical implications of this work by highlighting the impact of reifying self-fulfilling prophecies and propose a reorientation toward actions over accuracy as a priority for AI integration.

The desire to predict the future is related to a clinician's mandate to balance the intended benefit of medical interventions against possible risks. Outcome-prediction models (OPMs) utilize a patient's individual features to assist decision-making, offering a personalized approach to estimating a patient's outcome as a means of balancing benefits and risks. Artificial intelligence (AI) or machine learning (ML) approaches can perform these tasks dynamically, in real time, to support clinical decision-making. The accuracy of AI tools is widely considered to be the "prime directive," in that the main focus of translation efforts is to develop highly accurate models. The assumption is that an accurate model will inevitably yield better outcomes for patients, but what if that assumption is incorrect?

In a new paper, Van Amsterdam et al.[1] demonstrate how harmful decisions can result from accurate, well-calibrated models. Their work describes a set of cases wherein an OPM induces a care approach that results in worse outcomes for some patients. Importantly, they also show how the outcomes across this set of cases are not set up to be detected by dominant approaches to AI monitoring and evaluation. They highlight how the OPM induces a "self-fulfilling prophecy"—a prediction that brings about its own fulfillment.[2] In this situation, the personalized predictions made by the OPM influence care patterns in a way that reinforces the "correctness" of the original OPM predictions. For a subset of patients, this new pattern of "transformative self-fulfilling prophecies" can worsen their outcomes by inducing a treatment path that, without the model's prediction, would not have been pursued and thus would not have caused harm.[2] This provides a bracing reminder of the limitations of utilizing historical data for model training.

The self-fulfilling prophecy in medicine

Clinicians and clinical researchers have observed the epistemic and empirical problems of self-fulfilling prophecies in contexts such as the intensive-care unit, the transplant allocation of solid organs, and neuroprognostication.[3–5] In the intensive-care context, signs of a poor prognosis motivate decisions to limit interventions. The patient's subsequent deterioration appears to confirm that the initial decision was "correct" because the deterioration is interpreted as a function of the underlying disease (which, generally, is true). However, in a subset of cases, beliefs and attitudes toward disability, social functioning, race, and other non-causal factors can influence the impression that a clinician has about a patient's prognosis. These factors can then influence the apparent patterns of survivability after organ transplantation.[6] In these cases, the decision to withdraw or limit care interventions could give rise to an outcome that otherwise might have been prevented had a different decision been made.

Examples of this conundrum are found in the neonatal intensive-care unit in the treatment of infants with genetic conditions (e.g., Down syndrome and trisomy 18) and medical conditions such as hypoplastic left heart syndrome.[7] Historically, it was believed that these syndromes carried devastatingly poor outcomes, and therefore, attempts to prolong an infant's life were unethical. Accordingly, life-sustaining treatments were not typically offered. This pattern was subsequently captured in data, showing that infants with these syndromes had poor outcomes.[3] An accurate OPM based on these historical patterns would systematically predict high risk of mortality, persuading the clinician then to limit interventions, causing a self-fulfilling prophecy.

The disability community advocated for modifications in care delivery, and a more nuanced outcome picture has emerged. People with genetic differences such as Down syndrome and trisomy 18 can live rich lives, characterized by challenges not dissimilar to those of non-disabled people.[7] Clinicians have modified how they treat these patients in response to these observations, changing the pattern of outcomes to reveal that some conditions previously believed to be universally fatal or associated with unacceptable quality of life are not. The paradigm shift is not about moving toward one outcome or the other; the importance is in offering a more epistemically humble approach to counseling families where the recognition of our biases is accounted for in clinical decision-making.[8]

Outcome prediction and AI: Is accuracy misguided?

It is possible that when it comes to outcome predictions, we are thinking about AI in medicine all wrong. Although we prioritize predictive accuracy, we forget that many times more accurate predictions based on historical data are actually not better information to drive a contemporary decision.[9] For example, a recent systematic review of risk-stratification models revealed that only one-third of validated models offer a benefit, whereas the remainder have no positive influence or might even increase unnecessary healthcare utilization.[10] Better predictions are meaningless unless they facilitate better *decisions*.

Consider sepsis as an example. You can have an incredibly accurate algorithm that identifies cases of sepsis once treatment for sepsis has already been initiated. The value of the tool is in the timing of the opportunity to intervene: too early and you overtreat (and potentially cause harm); too late, and the model has no clinical utility. It is the value of the intervention—not the model accuracy—that provides the benefit.

Work by one of our authors highlights the importance of the "intervention ensemble" as a valuable concept for evaluating AI tools that goes beyond algorithmic validation.[11] The intervention ensemble as a concept reflects the notion that a model is meaningless without the surrounding practices that, together, sustain its value. For example, a clinically valuable prediction of sepsis would involve the identification of a specific point in the clinical workflow, a known set of actions recommended on the basis of particular outputs, and knowledge of which patient-level factors are well represented (e.g., comorbid conditions, genetic conditions, and medical complexities).

Van Amsterdam et al.'s results show how the quick move from empiric validation to deployment can induce harm by relying on model accuracy alone as a metric of benefit or clinical utility. By contrast, a silent trial would provide an opportunity to prospectively but non-interventionally test the model performance without affecting patient care.[12] This window can be used for testing the potential for self-fulfilling prophecies by facilitating the evaluation of the effects of counterfactuals, such as what would have happened had the clinician acted upon the prediction at a specified point in time. The clinical and developer teams can then identify the education needed for a clinical user to specify the actions that should and should not be taken on the basis of the prediction. For example, if the decision might induce a self-fulfilling prophecy, the clinician will need to consider multiple forms of evidence, consult with colleagues, and take time to ensure that the available evidence supports this decision.

Van Amsterdam et al. affirm the importance of randomized controlled trials (RCTs) to verify the prospective clinical value of an OPM to ensure that care is improved with the model rather than worsened. An RCT that evaluates the use of the OPM against its non-use (e.g., clinician intuition or a rudimentary prognostic calculation) would provide definitive information as to the relative merits of the OPM itself as a decision support tool, particularly given that outcome predictions might affect a range of care decisions by changing the clinical posture to the problem by modifying what seems "reasonable" in the context of the expected outcome. Notwithstanding the definitive knowledge gain here, we must attend to the relevant research ethics concepts that apply. When a model's outputs might influence the withdrawal or withholding of life-sustaining intervention, we must assess the risk to participants' welfare against the prospect of providing socially valuable information—essentially, whether or not equipoise exists.[13]

Additionally, it is often underappreciated that the same outcome prediction can be acted upon differently for different care goals. For example, predicting that a patient is at risk of deterioration can warrant an all-hands-on-deck approach to intervention or the transition to palliative care. The outcome is mortality for the latter, but that would not be considered a "bad" outcome if it was consistent with the patient's (or surrogate's) care goals.[8] This situation again calls for us to remember that the accuracy of a given prediction is not what determines the "goodness" of the model; it is our decisions that drive care and our decisions that can be good even in the most uncertain and ambiguous of circumstances.

Conclusion

Van Amsterdam et al.'s results push us to rethink what we assume is the value proposition of AI in medicine by forcing us to rethink what a "highly accurate" prediction in the context of an OPM actually means. The notion that we can do harm even with very accurate models is alarming but is not without precedent in the medical literature. By shifting focus from the model to our actions, we generate relevant knowledge about a model's utility and re-center the ethical impact on the patient as AI's prime directive. Additionally, by focusing on our actions, we can potentially mitigate an inherent risk of data-driven approaches to delivering care—the risk that medical science, by preserving the status quo through algorithmic delivery, is "stuck in amber." Algorithms codify and preserve the past, but decisions imagine new futures. With our actions as the guidepost, we can use data-driven insights without reifying history for accuracy's sake.

ACKNOWLEDGMENTS

M.D.M. acknowledges salary and research funding from the Hospital Research Foundation Group.

AUTHOR CONTRIBUTIONS

All authors contributed intellectually to the development of ideas, drafting, and editing of this manuscript.

DECLARATION OF INTERESTS

M.D.M. is a member of the *Patterns* advisory board.

REFERENCES

1. Van Amsterdam, W.A.C., van Geloven, N., Krijthe, J.H., Ranganath, R., and Cinà, G. (2025). When accurate prediction models yield harmful self-fulfilling prophecies. Patterns *6*, 101229. https://doi.org/10.1016/j.patter.2025.101229.
2. Mertens, M., King, O.C., van Putten, M.J.A.M., and Boenink, M. (2022). Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation. J. Med. Ethics *48*, 922–928. https://doi.org/10.1136/medethics-2020-106636.
3. McCradden, M.D., Anderson, J.A., and Cusimano, M.D. (2019). When is death in a child's best interest? Examining decisions following severe brain injury. JAMA Pediatr. *173*, 213–214. https://doi.org/10.1001/jamapediatrics.2018.4592.
4. Geocadin, R.G., Peberdy, M.A., and Lazar, R.M. (2012). Poor survival after cardiac arrest resuscitation: a self-fulfilling prophecy or biologic destiny? Crit. Care Med. *40*, 979–980. https://doi.org/10.1097/CCM.0b013e3182410146.
5. Wilkinson, D. (2009). The self-fulfilling prophecy in intensive care. Theor. Med. Bioeth. *30*, 401–410. https://doi.org/10.1007/s11017-009-9120-6.

6. National Academies of Sciences, Engineering, and Medicine (2022). Confronting and eliminating inequities in the organ transplantation system. In Realizing the Promise of Equity in the Organ Transplantation System, M. Hackmann, R.A. English, and K.W. Kizer, eds. (National Academies Press).

7. Kukora, S., Firn, J., Laventhal, N., Vercler, C., Moore, B., and Lantos, J.D. (2019). Infant with trisomy 18 and hypoplastic left heart syndrome. Pediatrics *143*, e20183779. https://doi.org/10.1542/peds.2018-3779.

8. Mazwi, M.L., Henner, N., and Kirsch, R. (2017). The role of palliative care in critical congenital heart disease. Semin. Perinatol. *41*, 128–132. https://doi.org/10.1053/j.semperi.2016.11.006.

9. Hunter, D.J. (2016). Uncertainty in the era of precision medicine. N. Engl. J. Med. *375*, 711–713. https://doi.org/10.1056/NEJMp1608282.

10. Oddy, C., Zhang, J., Morley, J., and Ashrafian, H. (2024). Promising algorithms to perilous applications: a systematic review of risk stratification tools for predicting healthcare utilisation. BMJ Health Care Inform. *31*, e101065. https://doi.org/10.1136/bmjhci-2024-101065.

11. McCradden, M.D., Joshi, S., Anderson, J.A., and London, A.J. (2023). A normative framework for artificial intelligence as a sociotechnical system in healthcare. Patterns *4*, 100864. https://doi.org/10.1016/j.patter.2023.100864.

12. McCradden, M.D., London, A.J., Gichoya, J.W., Sendak, M., Erdman, L., Stedman, I., Oakden-Rayner, L., Akrout, I., Anderson, J.A., Farmer, L.A., et al. (2025). CANAIRI: the Collaboration for Translational Artificial Intelligence Trials in healthcare. Nat. Med. *31*, 9–11. https://doi.org/10.1038/s41591-024-03364-1.

13. London, A.J. (2020). In Equipoise: integrating social value and equal respect in research with humans, A.S. Iltis and D. MacKay, eds. (Oxford University Press). https://doi.org/10.1093/oxfordhb/9780190947750.013.13.