Study Design Considerations for Post-Deployment Monitoring of AI/ML Models

Momin M. Malik, Ph.D.[1], Dave Watson, Ph.D.[2], Madison J. Beenken, M.S.[2], Shauna M. Overgaard, Ph.D.[3], Hanyin Wang, M.D.[4], Imad Absah, M.D.[1], Chung Il Wi, M.D.[1], Young J. Juhn, M.D., M.P.H.[1]

[1] Department of Pediatric and Adolescent Medicine, Mayo Clinic; [2] Department of Quantitative Health Sciences, Mayo Clinic; [3] Center for Digital Health, Mayo Clinic; [4] University of Illinois Urbana-Champaign

ABSTRACT

BACKGROUND

Classification models in artificial intelligence/machine learning (AI/ML), used for detection or prediction, are based on finding optimal correlations in a specific set of data. This often does a poor job of capturing underlying causal mechanisms, which makes such models fragile and potentially not generalize across conditions, settings, patient populations, potential interventions, and even across time for the same patient population. Thus, to safely and effectively use AI/ML models in clinical practice, we should robustly monitor their performance after deployment to be vigilant to possible degradations in performance and fairness. Yet there is little guidance about how to do such monitoring.

OBJECTIVE

Going beyond standard AI/ML development and evaluation study designs, we consider what specific study designs are needed for post-deployment monitoring.

METHODS

We make original linkages between machine learning methodology and biostatistics literature, particularly to diagnostic testing, for study design guidance on post-deployment model performance and fairness assessment.

RESULTS

We have two main findings.

1. For detection settings (e.g., screening and diagnosis), where an AI/ML model is used analogously to a diagnostic test, only looking at true positives among predicted positives to get positive predictive value is not sufficient. Without looking at how the model performs within predicted negatives, we cannot calculate accuracy, sensitivity, AUC, negative predictive value, nor calibration.
2. For prediction settings (e.g., prognosis), since AI/ML models do *not* take into account possibly successful interventions, models may seem to drift when they are clinically successful. Monitoring should take this into account.

The most direct way to address these issues is, respectively, random confirmatory testing on predicted negatives for detection, and maintaining long term "silent evaluation" sets for prediction. The costs of these will need to be weighed against the costs of not knowing all dimensions of performance or fairness of an AI/ML model.

CONCLUSIONS

There are specific study design considerations for post-deployment monitoring. Naïvely carrying out such monitoring might inaccurately measure model performance over time, or it may even lead to thinking the model is degrading exactly when it is working. Addressing these possible errors, and exploring other considerations, will be important when moving forward to robustly monitor the performance and fairness of AI/ML models in clinical deployment.

OBJECTIVES

We take two main settings: detection, and prediction.

Detection

- There is a condition that requires some onerous and/or expensive gold standard test to diagnose
- We get existing confirmed positives and negatives from the gold standard test (a case-control sample)
- Using that as the output variable, we build an AI/ML model using readily-available data (diagnosis codes, lab values, etc.)
- We apply the AI/ML to a general patient population
- We give the gold standard test to patients who the model scores as likely positive

Prediction

- There is some undesirable outcome
- Using previously realized cases, we build an AI/ML model with data that is available before the outcome happens
- We use this model to flag patients at risk, and potentially intervene to prevent the outcome

METHODS

Detection

- The detection setting is directly analogous to diagnostic testing,[1] so we can transfer lessons.
- Diagnostic tests are only valid for patients who are already positive on some "index" test. So, we should keep in mind that any AI/ML model fit to case-control data will only be valid for patients who would have been subjected to a screening tests anyway, and not the general population.
- Diagnostic tests made from case-control data can only give sensitivity and sensitivity, and not positive predictive value (PPV, also called precision) nor negative predictive value (NPV). PPV and NPV are clinically relevant, and will determine how useful an ML model is (low PPV means lots of false positives, leading to alert fatigue). So, we can't know PPV or NPV from development data.
- If we run a trial where we only perform confirmatory testing on predicted positives, we will *only* be able to calculate PPV, and not sensitivity or specificity in an application setting, nor AUC or NPV.

Prediction

- The goal is seldom passive prediction, but *intervention* to *prevent* bad outcomes. But "counterfactual prediction" (predicting results of an intervention) involves causality, and is *not* what usual AI/ML do.
- Doing causal modeling is hard; more feasible is finding ways to productively use passive predictions. But, we have to keep that gap in mind and manage the mismatch.
- As one example of the mismatch, Carter et al.[2] point out a paradoxical pattern: seemingly decreasing model performance ("drift") may be a sign of clinical success in acting on the passive predictions, and having prevented bad outcomes!

Exploring the problems articulated by these different literatures, we create recommendations for post-deployment monitoring.

RESULTS

Detection

If follow-up testing is done only for patients flagged by an AI/ML model, we only know the True Positives among the Predicted Positives, which only lets us calculate the PPV. We *cannot* calculate:

- sensitivity, (requires actual positives, including among predicted negatives);
- sensitivity (requires actual negatives and true negatives);
- negative predicted value (requires true negatives);
- AUC, accuracy, or calibration;
- fairness like risk ratios of False Omission Rate[3] or differential calibration.[4]

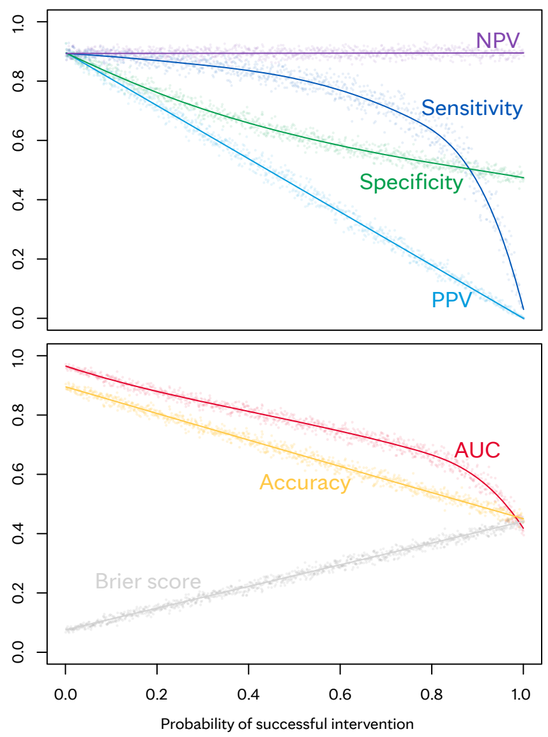
The most straightforward solution is testing of at least some predicted negatives.

Prediction

We use simulations to investigate the relationship between successful interventions and various metrics, an issue raised by Carter et al.[2]

All metrics show strong functional forms in increasing probability of successful intervention. While we do not work out what these functional forms are specifically, smoothing splines fit to the simulated data lets us see the overall pattern (Fig. 1). Only the Negative Predictive Value remains unchanged; all other metrics get worse. Accuracy goes to below 0.5 (worse than random, as we simulated balanced classes), as do Brier score and AUC (for which 0.5 is worse than random whether classes are balanced or not).

FIGURE 1: Performance shifts under intervention



For a fixed classifier, under simulated increasing probability of successful intervention on predicted positives (i.e., successfully changing true positives to false positives, which also changes relative counts of positives and negatives) on new data simulated from the same process on which the classifier was fit, all metrics other than NPV get worse.

DISCUSSION

These problems could be addressed by:

Detection: do random testing of predicted negatives, or something analogous.

- We can use power calculations to determine the number of predicted negatives to test.
- Potentially weigh the costs of such testing against the costs of not having insight into the performance and fairness of an AI/ML classification model.

Prediction: have a background "silent evaluation" to measure model performance without the confounder of potentially successful interventions.

- This set might also allow calculation of an average treatment effect from actions taken in response to model outputs.
- Again, power calculations may be necessary to determine a sufficient sample size for such a silent evaluation set.
- Again, the harms of patients not benefiting from a potentially assistive intervention will need to be weighed against not having insight into performance/fairness.
- Maybe the relationships estimated in Fig. 1 can be applied as a **corrective** under an estimated probability of successful intervention, but **we do not recommend this** as it would treat the probability of successful intervention as homogeneous, which is not realistic. But, further modeling may find ways to create corrections.

If these straightforward solutions are too costly, perhaps further exploration can find other ways of addressing the challenges we articulate.

CONCLUSIONS

Post-deployment monitoring is important, but not straightforward. There are under-examined and poorly understood relationships between machine learning model performance and actions taken on their basis. We provide theory-based initial guidance and call for further investigation, including empirical study.

ACKNOWLEDGMENTS

This work was funded by Mayo Clinic Health System Research, and the Mayo Clinic Children's AI Program.

REFERENCES

1. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. John Wiley & Sons, Inc.; 2011.
2. Carter RE, Anand V, Harmon DM Jr, Pellikka PA. Model drift: when it can be a sign of success and when it can be an occult problem. Intell Based Med. 2022;6:100058. doi: 10.1016/j.ibmed.2022.100058.
3. Rodolfa K, Saleiro P, Ghani R. Bias and fairness. In: Foster I, Ghani R, Jarmin RS, Kreuter F, Lane J, editors. Big data and social science. 2nd ed. Chapman and Hall/CRC; 2021. p. 281-312. doi: 10.1201/9780429324383-11.
4. Davis SE, Dorn C, Park DJ, Matheny ME. Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. J Am Med Inform Assoc. 2025 May 1;32(5):845-854. doi: 10.1093/jamia/ocaf039.