Change in machine learning model performance upon retraining after deployment into clinical practice: The real-world effect of model predictions on clinician actions, outcome labels, and the potential for contamination bias

Michael Colacci MD PhD[1-3], George-Alexandru Adam PhD[4], Chloe Pou-Prom MSc[1], Anna Goldenberg PhD[4], Amol Verma MD, MPhil[1-3], Muhammad Mamdani PharmD, MA, MPH[1-3]

[1] Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, Canada, [2] Department of Medicine, University of Toronto, Toronto, Canada
[3] Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada [4] Department of Computer Science, University of Toronto, Toronto, Canada

BACKGROUND

CHARTWatch is an AI model that was implemented at St. Michael's Hospital in October 2020 and predicts inpatient deterioration on medical wards.[1]

AI Models may degrade over time due to factors like data drift and may require retraining. However, retraining a deployed model using post-deployment data may worsen performance due to **contamination bias**—a phenomenon where the model changes outcomes it later uses to retrain, as shown in recent simulation studies.[2-5]

To inform whether to retrain CHARTwatch, we sought to **quantify the degree of contamination bias** and explore strategies to mitigate it.

METHODS

Part 1: Prospective cohort study

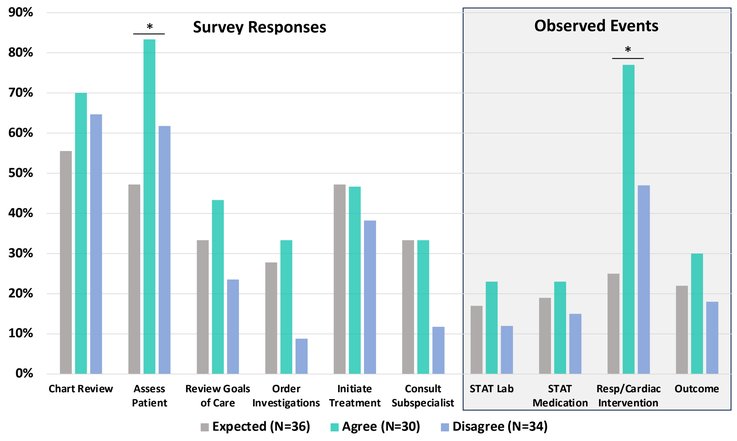
- Evaluate how often CHARTwatch prevents patient deterioration (effective intervention rate) through real-time surveys and chart review for 100 consecutive alerts
- Quantify the overall outcome frequency and model recall

Part 2: Model retraining evaluations

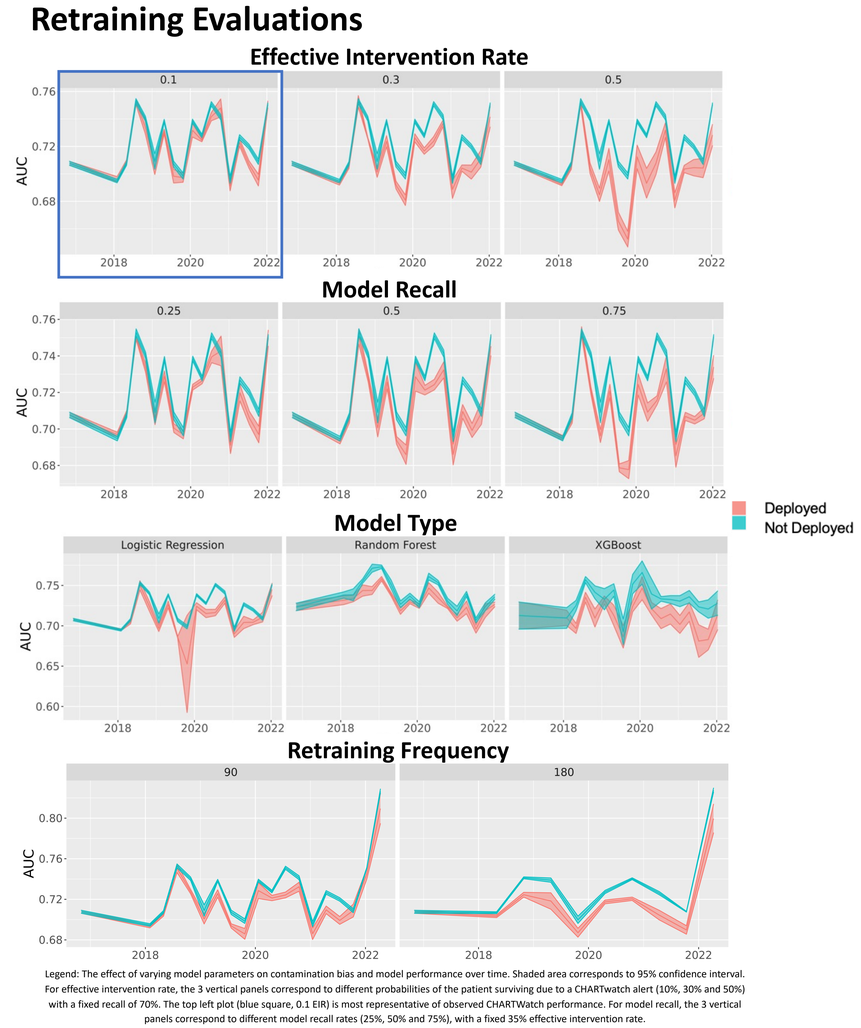
- Evaluate the change in model performance due to contamination bias when retraining with a range of effective intervention rates , recall frequencies, and ML model types, with values informed by the prospective cohort study
- Contamination bias mitigation: Evaluate the effect of removing potentially confounded outcomes prior to retraining on the magnitude of contamination bias and overall model performance

RESULTS

Prospective Cohort Study

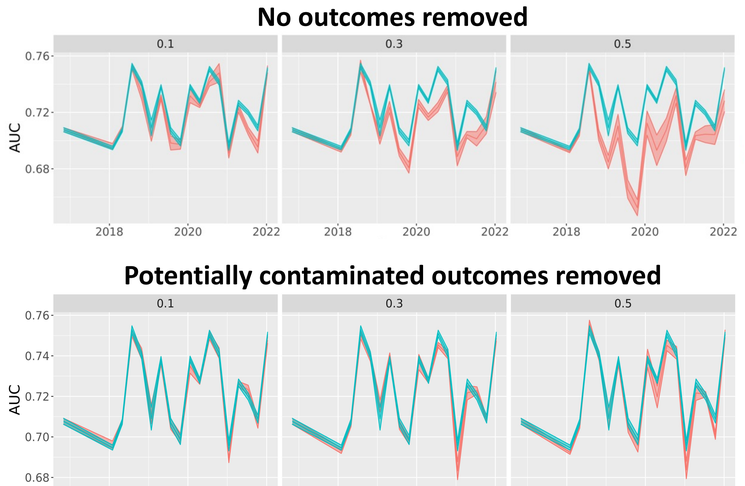


Retraining Evaluations



Legend: The effect of varying model parameters on contamination bias and model performance over time. Shaded area corresponds to 95% confidence interval. For effective intervention rate, the 3 vertical panels correspond to different probabilities of the patient surviving due to a CHARTwatch alert (10%, 30% and 50%) with a fixed recall of 70%. The top left plot (blue square, 0.1 EIR) is most representative of observed CHARTWatch performance. For model recall, the 3 vertical panels correspond to different model recall rates (25%, 50% and 75%), with a fixed 35% effective intervention rate.

Contamination Bias Mitigation



Legend: The effect of removing potentially contaminated outcomes before retraining on model performance over time. Shaded area corresponds to 95% confidence interval. In the top panel, all patient outcomes were retained in model retraining. In the bottom panel, for the deployed model, all patients who had a CHARTwatch alert but did not experience an outcome were removed from retraining as they were potentially contaminated. The three vertical panels correspond to model effective intervention rates of 10%, 30% and 50%.

DISCUSSION

Clinician agreement with CHARTwatch predictions is associated with downstream clinical actions.

We present a framework for using clinician perception/actions and model parameters to estimate contamination bias.

For CHARTwatch, contamination bias at any retraining interval is limited (△AUC <2%) but the impact over time is summative.

Removing potentially confounded outcomes may help mitigate contamination bias during retraining.

REFERENCES:

1. Verma AA, Stukel TA, Colacci M, et al. Clinical evaluation of a machine learning-based early warning system for patient deterioration. *CMAJ*. 2024;196(30):E1027-E1037. doi:10.1503/cmaj.240132
2. Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. Hidden Risks of Machine Learning Applied to Healthcare: Unintended Feedback Loops Between Models and Future Data Causing Model Degradation. In: *Proceedings of the 5th Machine Learning for Healthcare Conference*; 2020.
3. Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. Error Amplification When Updating Deployed Machine Learning Models. In: *Proceedings of Machine Learning Research*. Vol 182. ; 2022.
4. Vaid A, Sawant A, Suarez-Farinas M, et al. Implications of the Use of Artificial Intelligence Predictive Models in Health Care Settings : A Simulation Study. *Ann Intern Med*. 2023;176(10):1358-1369. doi:10.7326/M23-0949
5. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*. 2021;385(3). doi:10.1056/nejmc2104626
6. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless... *Journal of the American Medical Informatics Association*. 2019;26(12). doi:10.1093/jamia/ocz145

UNITY HEALTH TORONTO
Dalla Lana School of Public Health
CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada
Eliot Phillipson CLINICIAN-SCIENTIST TRAINING PROGRAM