Clinical evaluation of a machine learning–based early warning system for patient deterioration

Amol A. Verma MD MPhil, Therese A. Stukel PhD, Michael Colacci MD, Shirley Bell RN, Jonathan Ailon MD, Jan O. Friedrich MD DPhil, Joshua Murray MSc, Sebnem Kuzulugil PhD, Zhen Yang MSc, Yuna Lee MD MEd, Chloe Pou-Prom MSc, Muhammad Mamdani PharmD

■ Cite as: *CMAJ* 2024 September 16;196:E1027-37. doi: 10.1503/cmaj.240132

Abstract

Background: The implementation and clinical impact of machine learning–based early warning systems for patient deterioration in hospitals have not been well described. We sought to describe the implementation and evaluation of a multifaceted, real-time, machine learning–based early warning system for patient deterioration used in the general internal medicine (GIM) unit of an academic medical centre.

Methods: In this nonrandomized, controlled study, we evaluated the association between the implementation of a machine learning–based early warning system and clinical outcomes. We used propensity score–based overlap weighting to compare patients in the GIM unit during the intervention period (Nov. 1, 2020, to June 1, 2022) to those admitted during the pre-intervention period (Nov. 1, 2016, to June 1, 2020). In a difference-in-differences analysis, we compared patients in the GIM unit with those in the cardiology, respirology, and nephrology units who did not receive the intervention. We retrospectively calculated system predictions for each patient in the control cohorts, although alerts were sent to clinicians only during the intervention period for patients in GIM. The primary outcome was non-palliative in-hospital death.

Results: The study included 13 649 patient admissions in GIM and 8470 patient admissions in subspecialty units. Non-palliative deaths were significantly lower in the intervention period than the pre-intervention period among patients in GIM (1.6% v. 2.1%; adjusted relative risk [RR] 0.74, 95% confidence interval [CI] 0.55–1.00) but not in the subspecialty cohorts (1.9% v. 2.1%; adjusted RR 0.89, 95% CI 0.63–1.28). Among high-risk patients in GIM for whom the system triggered at least 1 alert, the proportion of non-palliative deaths was 7.1% in the intervention period, compared with 10.3% in the pre-intervention period (adjusted RR 0.69, 95% CI 0.46-1.02), with no meaningful difference in subspecialty cohorts (10.4% v. 10.6%; adjusted RR 0.98, 95% CI 0.60–1.59). In the difference-in-differences analysis, the adjusted relative risk reduction for non-palliative death in GIM was 0.79 (95% CI 0.50–1.24).

Interpretation: Implementing a machine learning–based early warning system in the GIM unit was associated with lower risk of non-palliative death than in the pre-intervention period. Machine learning–based early warning systems are promising technologies for improving clinical outcomes.

Predicting, preventing, and rapidly responding to patient deterioration in hospital is a major goal for improving patient safety. Unrecognized clinical deterioration is the leading cause of unplanned transfers to intensive care units (ICUs) in hospital, which are associated with longer hospital stays and higher mortality than ICU transfers that occur directly from the emergency department.[1–3] Researchers have tried to predict in-hospital deterioration,[4] including using complex statistical models[5] and machine learning methods[6,7]

Evidence about the effectiveness of using prediction tools as early warning systems to detect and reduce patient deterioration in hospitals is mixed, despite widespread use of such tools.[8,9] Recent literature reviews[10,11] have identified only a small number of studies that report the effect of early warning systems on hospital or 30-day mortality, many of which had serious methodological limitations such as reporting only unadjusted or uncontrolled estimates. In 1 review,[10] of the 6 studies that used robust methodology, only a single study reported a significant improvement in patient outcomes with the implementation of an early warning system. This 19-hospital study at Kaiser Permanente Northern California found that an automated prediction model with remote nurse monitoring and on-the-ground intervention by rapid response teams was associated with a 16% relative reduction in 30-day mortality.[12] The technical and clinical features of statistically advanced early warning systems that may be associated with improvements in clinical outcomes remain uncertain.[11]

The purpose of this study was to explore the association between the implementation of a multifaceted, machine learning–based early warning system — composed of a real-time risk prediction tool, clinical alerts, and a clinical care pathway for high-risk patients — and clinical outcomes among patients in the general internal medicine (GIM) unit of an academic health centre.

Methods

Study design and setting

We conducted a nonrandomized, controlled study at St. Michael's Hospital, an inner-city academic health centre in Toronto, Canada. The GIM unit includes around 70 beds managed by 5 medical teams, including an attending physician, residents, medical students, and interprofessional staff. The GIM unit has a 4-bed step-up unit that can provide more intensive nursing, and the hospital has a critical care response team to provide ICU outreach services to deteriorating ward patients. The early warning system, called CHARTwatch, was implemented on the GIM unit in the fall of 2020 through a phased implementation between August and October, after 3 years of development and validation studies.[7,13–15] Before implementing CHARTwatch, the critical care response team was activated based on physician or nurse judgment, but no formal deterioration detection score was available to guide these decisions. The prespecified primary aim of CHARTwatch was to reduce non-palliative deaths (defined as deaths without receiving palliative care) by expediting interventions that might reduce risk of death and by prompting earlier consultation with palliative care specialists when clinicians felt it was appropriate.

This manuscript is reported in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis and Artificial Intelligence (TRIPOD+AI)[16] and Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence (DECIDE-AI) items.[17]

Data collection

We collected data from the hospital's electronic medical record and routine administrative data systems. Patient diagnoses were coded using the Canadian version of the *International Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10-CA), which are mandatorily coded at the time of hospital discharge and reported to the Canadian Institute for Health Information.

Participants

The program's evaluation was planned before the COVID-19 pandemic and was initially conceived as a matched cohort study comparing patients in GIM before and after deployment. Given that the pandemic could confound pre–post analysis, we decided to also measure changes in the pre- and post-intervention periods that occurred on other medical wards in the hospital that did not receive the intervention. We included the cardiology, nephrology, and respirology units as comparators because their patient populations are nonobstetrical, largely nonsurgical, and mostly admitted acutely through the emergency department; they have similar interprofessional team structures.

We included 4 patient cohorts admitted through the emergency department. The GIM intervention cohort included all patients cared for by a GIM team on the GIM unit between Nov. 1, 2020, and June 1, 2022. We ended the observation period at this time because the CHARTwatch machine learning model was changed to accommodate new software and data pipelines. The GIM pre-intervention cohort included all patients cared for by a GIM team on the GIM unit between Nov. 1, 2016, and June 1, 2020. The contemporaneous subspecialty cohort included all patients admitted to the cardiology, nephrology, or respirology units between Nov. 1, 2020, and June 1, 2022. The pre-intervention subspecialty cohort included all patients admitted to the cardiology, nephrology, or respirology units between Nov. 1, 2016, and June 1, 2020.

To facilitate statistical comparisons across the cohorts, we retrospectively calculated CHARTwatch predictions for each patient in the control cohorts, just as we would have if the intervention had been live. These predictions were not presented to clinicians and did not influence care, but allowed us to statistically balance cohorts based on patients' baseline risk of death in hospital and to identify the patients who would have received high-risk alerts in each of the control groups. We were then able to compare outcomes and processes of care among the subgroup of patients who were high-risk in the intervention group and control groups.

We excluded patients with COVID-19 (based on ICD-10 codes U07.1 or U07.2),[18] as they did not exist in the control period, and patients with influenza (ICD-10 codes J09, J10.0, J10.1, J10.8, J11.0, J11.1, or J11.8),[19] as nearly no influenza admissions occurred in the intervention period. Because the primary aim was a reduction in non-palliative deaths, we excluded patients who had a preadmission palliative care comorbidity (defined by the ICD-10 diagnosis code Z51.5) and patients whose code status was to receive comfort measures at baseline. Patients who received palliative care after admission remained in the study, as an aim of the intervention was to improve timely access to palliative care.

Intervention

The intervention consisted of a machine learning model that used real-time data from the electronic medical record to predict patient deterioration, communication of predictions to physicians and nurses, a clinical pathway for high-risk patients, and a multidisciplinary implementation team that monitored and refined the intervention regularly after it was launched.[7,13–15] Appendix 1, Sections 1–5, available at www.cmaj.ca/lookup/doi/10.1503/cmaj.240132/tab-related-content, provides a more detailed description of each component.

The deterioration prediction model was a time-aware multivariate adaptive regression spline (MARS) model (Appendix, Sections 1–4). The model is made time-aware by incorporating risk score predictions from earlier in the encounter, the change in risk score since the previous assessment, and summaries of changes in the risk score over time. To communicate risk to clinicians, patients were sorted into high-, medium-, and low-risk categories. A specific clinical care pathway was developed and implemented for high-risk patients, whereas medium- and low-risk categories were provided to inform clinicians about the patient's status with no specific care pathway. Based on clinical input, the threshold for the high-risk group was set at a positive predictive value (PPV) of 30% for deterioration during the hospital admission in the training data since 1 truly alerted patient for every 2 falsely alerted patients was deemed an acceptable number of false alarms. Based on this threshold, the model had a sensitivity of 53% and a PPV of 31% for detecting clinical deterioration during the hospital admission (death or transfer to the ICU, step-up care, or the palliative care unit) in the held-out testing data. Additional alarm silencing rules were enacted to minimize alarm fatigue (Appendix 1).

Model predictions were reported directly to physicians and charge nurses via text paging and emails, initially 3 times per day, and then increased to hourly on Jan. 19, 2021 (Appendix 1). Clinicians were asked to adhere to a clinical pathway for high-risk patients, including physician reassessment within 1 hour, increased monitoring of vital signs by nurses, and triggers for possible palliative care consultation when the most responsible physicians felt it was appropriate.[15] The GIM teams escalated care to the ICU outreach team as needed.

Outcomes

The primary outcome was non-palliative in-hospital deaths, defined as deaths that occurred without a documented palliative care intervention, identified by the presence of a palliative care ICD-10 code or transfer to the inpatient palliative care unit. Secondary outcomes included overall deaths, palliative deaths and transfers (defined as the composite of deaths with palliative care or transfers to the inpatient palliative care unit), transfer to ICU, the composite of transfer to ICU or death, and length of hospital stay. The first CHARTwatch prediction was calculated in all cohorts at the time of transfer from the emergency department to an inpatient ward. This time-point was set as the baseline, and we measured all outcomes from this time until hospital discharge. For the subgroup analysis of high-risk patients, we measured outcomes after the first high-risk prediction.

We also measured the following pre-specified processes of care in the 24 hours after the first high-risk CHARTwatch alert, selected based on clinical expertise, the published literature,[11,20] and data that could be reliably extracted from the hospital electronic medical record, namely new antibiotics, glucocorticoids, intravenous fluids, radiography, computed tomography, ultrasonography, magnetic resonance imaging, new physician order for code status, number of vital sign measurements documented in the medical record, and transfer to the GIM step-up unit.

Covariates

We included prespecified covariates that were expected to influence the risk of in-hospital death based on the clinical and scientific expertise of our team, namely admitting service (GIM, cardiology, nephrology, respirology), age, sex, CHARTwatch risk score at baseline (defined as the time of first transfer to a medical ward or, for the high-risk subgroup, the time of the first high-risk alert), number of hospital admissions in previous 6 months, number of admissions for the same primary diagnosis in the previous 6 months,[21] admission on a weekend,[22] Charlson comorbidities,[23] calendar month at admission, most responsible discharge diagnosis (categorized using the Clinical Classifications Software Revised),[24,25] vital signs, code status at baseline (categorized as ICU acceptable, do not resuscitate or not for ICU, and not documented), homelessness, neighbourhood material resources and neighbourhood racialized and newcomer population (as defined by the Ontario Marginalization Index, categorized into quintiles),[26] and ICU admission before transfer to a GIM or subspecialty ward. Race, ethnicity, and language data were not collected reliably for individual patients at our hospital.[27]

Statistical analysis

We reported patient characteristics for each of the 4 study cohorts. The primary comparison was between patients admitted to the GIM unit where CHARTwatch was deployed during the intervention period and those admitted to this unit during the pre-intervention period. We reported changes during the same periods in medical subspecialty units where CHARTwatch was not deployed. We then stratified the analysis into high-risk and low-risk groups. We categorized patients as high risk if they received at least 1 high-risk CHARTwatch prediction. Because there was no specific intervention for medium- or low-risk patients, we analyzed these patients together as the low-risk group for this study. We expected the intervention to affect the high-risk patient group, as this was the focus of the high-risk clinical care pathway. Low-risk patients received standard care, and the results in this group served as a balancing measure to identify unintended consequences of diverting resources toward high-risk patients.

We used propensity score–based overlap weights to balance differences in measured baseline characteristics that could influence the risk of in-hospital death. The propensity score, defined as the probability of being admitted during the intervention period versus the control period, was estimated using multivariable logistic regression separately for the GIM and subspecialty cohorts. We fit a separate propensity score model for the high-risk subgroups. The general model included covariates measured at baseline before the first CHARTwatch prediction; the high-risk model was the same but the CHARTwatch risk score, vital signs, and code status were measured at the time of, or immediately before the first high-risk prediction. We modelled vital signs as restricted cubic splines with 3 knots.

Each patient was weighted according to the overlap weight,[28] which is the probability of being assigned to the opposite exposure group based on propensity score. Patients with a high propensity score, that is, a high probability of receiving either exposure, were assigned the largest weights, thus reducing the influence of observations at the extremes of the probability distribution. Overlap weights produce the smallest standard errors among weight-balancing approaches and achieve perfect balance for covariates included in the propensity score.[28,29]

We used Poisson regression, weighting by overlap weights, to compare binary outcomes, and linear models to compare continuous outcomes (i.e., length of stay). We included a robust variance estimator to account for weighting. Follow-up was censored at 90 days post-admission if patients did not experience an outcome and were not discharged.

We performed a difference-in-differences analysis[30] as a secondary analysis to compare the observed changes on GIM versus subspecialty wards. This allowed us to control for secular trends and account for the known, stable differences between patients in GIM and subspecialty wards (e.g., some patients in cardiology units are admitted for cardiac procedures, the nephrology unit includes patients admitted for kidney transplants). We combined the GIM and subspecialty cohorts and created a propensity score model to weight the patients. The propensity score modelled the probability of being admitted during the intervention versus control periods and included the same covariates as the overall model, as well as type of admission (GIM v. subspecialty).

We performed analyses using R version 3.6.3 and the tidyverse,[31] tableone,[32] PSweight,[33] survey,[34] and geepack[35] packages.

Ethics approval

Ethics approval was obtained from the St. Michael's Hospital Research Ethics Board (no. 19-317).

Results

Patient characteristics

Study cohorts are described in Figure 1, and patient characteristics are shown in Table 1 and Appendix 1, Supplemental Tables 1–3. The study included 13 649 patients admitted to GIM (9626 pre-intervention and 4023 intervention) and 8470 patients admitted to subspecialty units (6103 pre-intervention and 2367 contemporaneous). In GIM, 482 patients became high risk during the intervention period and 1656 patients became high risk in the control period.

In the GIM intervention cohort, the median age was 68 (interquartile range [IQR] 55–80) years, 43.3% were female, 10.5% were not for ICU-level care, 14.5% were experiencing homelessness, 25.9% were living in neighbourhoods with the lowest material resources, and 34.7% were living in neighbourhoods with the greatest racialized and newcomer populations. In the subspecialty contemporaneous cohort, the median age was 66 (IQR 54–77) years, 41.1% were female, 4.5% were not for ICU-level care, 3.8% were experiencing homelessness, 25.6% were living in neighbourhoods with the lowest material resources, and 36.1% were living in neighbourhoods with the greatest racialized and newcomer populations (Appendix 1, Table S1).

Before weighting, baseline characteristics, except for median age, were generally well balanced; after weighting, measured risk factors were perfectly balanced between the groups.

Outcomes in the GIM and subspecialty cohorts

Outcomes are reported in Table 2 and Figure 2. After weighting, non-palliative deaths were significantly lower in the GIM intervention group (2.1%) than in the GIM pre-intervention group (1.6% intervention; adjusted relative risk [RR] 0.74, 95% confidence interval [CI] 0.55–1.00). We did not observe a significant difference in overall deaths (3.7% pre-intervention v. 3.4% intervention; adjusted RR 0.93, 95% CI 0.75–1.14), palliative deaths or transfers (2.0% pre-intervention v. 2.3% intervention; adjusted

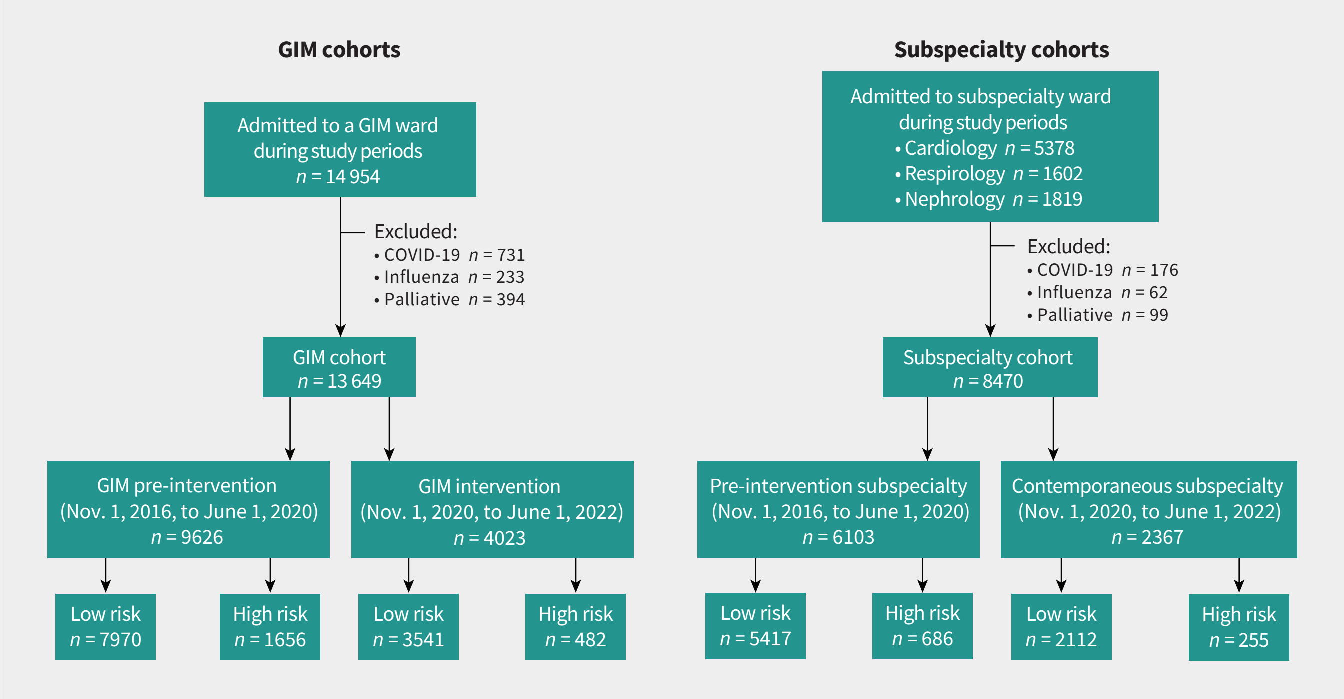
GIM cohorts
Admitted to a GIM ward during study periods n = 14 954
Excluded:
• COVID-19 n = 731
• Influenza n = 233
• Palliative n = 394
GIM cohort n = 13 649
GIM pre-intervention (Nov. 1, 2016, to June 1, 2020) n = 9626
GIM intervention (Nov. 1, 2020, to June 1, 2022) n = 4023
Low risk n = 7970
High risk n = 1656
Low risk n = 3541
High risk n = 482
Subspecialty cohorts
Admitted to subspecialty ward during study periods
• Cardiology n = 5378
• Respirology n = 1602
• Nephrology n = 1819
Excluded:
• COVID-19 n = 176
• Influenza n = 62
• Palliative n = 99
Subspecialty cohort n = 8470
Pre-intervention subspecialty (Nov. 1, 2016, to June 1, 2020) n = 6103
Contemporaneous subspecialty (Nov. 1, 2020, to June 1, 2022) n = 2367
Low risk n = 5417
High risk n = 686
Low risk n = 2112
High risk n = 255

Figure 1: Study cohorts. Patients were categorized as high risk if they received at least 1 high-risk alert in the intervention cohort or if they would have received one had the intervention been active in the control cohorts. All other patients were defined as low risk. See Related Content tab for accessible version. Note: GIM = general internal medicine.

Table 1 (part 1 of 2): Patient characteristics of general internal medicine (GIM) cohorts

	Unweighted			Weighted		
Characteristic	No. (%) of patients in GIM pre-intervention* *n* = 9626	No. (%) of patients in GIM intervention* *n* = 4023	SMD	Percentage of patients in GIM pre-intervention* *n* = 9626	Percentage of patients in GIM intervention* *n* = 4023	SMD†
Age, yr, median (IQR)	66.1 (52.7–78.79)	68.3 (54.9–79.7)	0.085	67.3	67.7	0.008
Sex, female	3915 (40.7)	1741 (43.3)	0.053	42.3	42.3	0
First CHARTwatch score, mean ± SD	0.09 ± 0.02	0.09 ± 0.02	0.217	0.09	0.09	0
Admissions in previous 6 mo			0.073			0
0	8330 (86.5)	3569 (88.7)		88.0	88.0	
1	715 (7.4)	271 (6.7)		7.0	7.0	
≥ 2	581 (6.0)	183 (4.6)		5.0	5.0	
Admissions for same primary diagnosis in previous 6 mo			0.075			0
0	9200 (95.6)	3902 (97.0)		96.6	96.6	
1	293 (3.0)	82 (2.0)		2.4	2.4	
≥ 2	133 (1.4)	39 (1.0)		1.1	1.1	
Admitted on weekend	2184 (22.7)	1030 (25.6)	0.068	24.6	24.6	0
Preadmission comorbidities						
Myocardial infarction	66 (0.7)	19 (0.5)	0.028	0.5	0.5	0
Heart failure	897 (9.3)	323 (8.0)	0.046	8.5	8.5	0
Peripheral vascular disease	105 (1.1)	47 (1.2)	0.007	1.1	1.1	0
Cerebrovascular disease	716 (7.4)	429 (10.7)	0.113	9.8	9.8	0
Dementia	281 (2.9)	192 (4.8)	0.096	4.1	4.1	0
Chronic pulmonary disease	909 (9.4)	203 (5.1)	0.170	6.1	6.1	0
Connective tissue disease	129 (1.3)	35 (0.9)	0.045	1.0	1.0	0
Peptic ulcer disease	113 (1.2)	63 (1.6)	0.034	1.4	1.4	0
Liver disease, mild	365 (3.8)	192 (4.8)	0.048	4.3	4.3	0
Liver disease, moderate or severe	171 (1.8)	39 (1.0)	0.069	1.1	1.1	0
Diabetes	1351 (14.0)	556 (13.8)	0.006	14.1	14.1	0
Diabetes with complications	1834 (19.1)	727 (18.1)	0.025	18.5	18.5	0
Hemiplegia	145 (1.5)	76 (1.9)	0.030	1.8	1.8	0
Kidney disease	455 (4.7)	226 (5.6)	0.040	5.3	5.3	0
Cancer	484 (5.0)	257 (6.4)	0.059	5.9	5.9	0
Cancer with metastasis	189 (2.0)	86 (2.1)	0.012	2.0	2.0	0
AIDS	79 (0.8)	36 (0.9)	0.008	0.9	0.9	0
Temperature, mean ± SD	36.4 ± 0.6	36.4 ± 0.6	0.030	36.4 ± 0.6	36.4 ± 0.6	0.011
Heart rate, mean ± SD	83.1 ± 14.7	81.9 ± 14.4	0.086	82.2 ± 14.5	82.2 ± 14.5	0.001
Systolic blood pressure, mean ± SD	129.4 ± 20.2	131.2 ± 0.5	0.086	130.7 ± 20.4	130.7 ± 20.5	0.004
Diastolic blood pressure, mean ± SD	73.2 ± 10.9	74.1 ± 10.6	0.086	73.9 ± 10.8	73.9 ± 10.7	0.003
Oxygen saturation, mean ± SD	96.4 ± 2.4	96.8 ± 2.2	0.173	96.7 ± 2.3	96.7 ± 2.3	0.002
Respiratory rate, mean ± SD	19.1 ± 2.0	18.8 ± 1.8	0.164	18.9 ± 1.9	18.9 ± 1.9	0.01
Code status			0.191			0
Not for intensive care	856 (8.9)	421 (10.5)		10.1	10.1	
Intensive care acceptable	4331 (45.0)	2124 (52.8)		50.2	50.2	
Not documented	4439 (46.1)	1478 (36.7)		39.8	39.8	
Homelessness	1619 (16.8)	583 (14.5)	0.064	15.1	15.1	0

Table 1 (part 2 of 2): Patient characteristics of general internal medicine (GIM) cohorts

	Unweighted			Weighted		
Characteristic	No. (%) of patients in GIM pre-intervention* *n* = 9626	No. (%) of patients in GIM intervention* *n* = 4023	SMD	Percentage of patients in GIM pre-intervention* *n* = 9626	Percentage of patients in GIM intervention* *n* = 4023	SMD†
Neighbourhood material resources			0.104			0
Q1 (most resources)	2441 (25.4)	1118 (27.7)		27.2	27.2	
Q2	1229 (12.8)	533 (13.2)		13.1	13.1	
Q3	1152 (11.97)	514 (12.8)		12.6	12.6	
Q4	1323 (13.7)	589 (14.6)		14.5	14.5	
Q5 (least resources)	2934 (30.5)	1044 (25.9)		27.1	27.1	
Missing	547 (5.7)	225 (5.6)		5.5	5.5	
Neighbourhood racialized and newcomer populations			0.084			0
Q1 (least)	316 (3.3)	108 (2.7)		27.9	27.9	
Q2	501 (5.2)	225 (5.6)		5.5	5.5	
Q3	1397 (14.5)	688 (17.1)		16.4	16.4	
Q4	3507 (36.4)	1382 (34.4)		35.1	35.1	
Q5 (most)	3358 (34.9)	1395 (34.7)		34.7	34.7	
Missing	547 (5.7)	225 (5.6)		5.5	5.5	

Note: IQR = interquartile range, SD = standard deviation, SMD = standardized mean difference.
*Unless indicated otherwise.
†Standardized differences after overlap weighting based on propensity score are, by definition, 0. For clarity and simplicity, we elected to present some variables in this table more simply than they were inputted in the propensity score model (e.g., age was entered as 10-year age bands and vital signs were modelled using restricted cubic splines). Thus, we report the standardized differences for the summary statistic presented in this table, which is why the standardized difference for some variables is greater than 0.

RR 1.14, 95% CI 0.88–1.49), ICU transfers (3.6% pre-intervention v. 3.9% intervention, adjusted RR 1.09, 95% CI 0.89–1.32), or length of hospital stay.

In the subspecialty cohorts, after weighting, there was no significant difference in non-palliative deaths (2.1% pre-intervention v. 1.9% contemporaneous; adjusted RR 0.89, 95% CI 0.63–1.28), overall deaths, palliative deaths or transfers, ICU transfers, or length of hospital stay.

Outcomes and processes in the high-risk subgroup

Weighted and unweighted outcomes in the high-risk subgroup are reported in Figure 3 and Appendix 1, Table S4. After weighting, the proportion of non-palliative deaths in the pre-intervention GIM cohort was 10.3% and 7.1% in the intervention period (adjusted RR 0.69, 95% CI 0.46–1.02). We did not observe a significant difference in overall deaths (16.6% pre-intervention v. 15.4% intervention; adjusted RR 0.92, 95% CI 0.71–1.20), palliative deaths or transfers (7.5% pre-intervention v. 9.7% intervention; adjusted RR 1.30, 95% CI 0.89–1.87), ICU transfers (13.1% pre-intervention v. 15.1% intervention; adjusted RR 1.15, 95% CI 0.87–1.52), or length of hospital stay.

In the subspecialty cohorts, after weighting, there was no significant difference in non-palliative deaths (10.6% pre-intervention v. 10.4% contemporaneous; adjusted RR 0.98, 95% CI 0.60–1.59) or secondary outcomes.

Processes of care in the 24 hours after the first high-risk alert are reported in Table 3. After weighting, patients in the GIM intervention cohort were more likely than controls to receive antibiotics (28.9% pre-intervention v. 49.4% intervention, $p < 0.001$), to receive systemic glucocorticoids (10.4% pre-intervention v. 17.0% intervention, $p = 0.001$), and to have vital signs measured more frequently (median 3 [IQR 2–5] measurements pre-intervention v. median 5 [IQR 4–7] measurements intervention, $p < 0.001$). After weighting, patients in the contemporaneous subspecialty cohort were also more likely to receive antibiotics (34.9% pre-intervention v. 47.0% contemporaneous, $p = 0.004$).

We did not observe any changes in imaging use, code status orders, or intravenous fluid orders in either the GIM or subspecialty groups.

Outcomes among low-risk patients in GIM

Compared with the pre-intervention period, we did not observe any significant change in non-palliative deaths (adjusted RR 1.09, 95% CI 0.67–1.77) or overall death (adjusted RR 1.19, 95% CI 0.85–1.66) among low-risk patients in GIM after overlap weighting (Appendix 1, Table S5). This group did have an increase in ICU transfers (1.6% pre-intervention v. 2.2% intervention; adjusted RR 1.39, 95% CI 1.03–1.86).

Table 2: Clinical outcomes in the general internal medicine (GIM) and subspecialty cohorts

	Unweighted			Weighted		
Outcome*	**No. (%) of patients in pre-intervention†**	**No. (%) of patients in intervention†**	***p* value**	**Percentage of patients in pre-intervention†**	**Percentage of patients in intervention†**	**RR (95% CI)**
Primary outcome, GIM						
Non-palliative death	207 (2.2)	59 (1.5)	0.01	2.1	1.6	0.74 (0.55–1.00)
Primary outcome, subspecialty						
Non-palliative death	128 (2.1)	43 (1.8)	0.5	2.1	1.9	0.89 (0.63–1.28)
Secondary outcomes, GIM						
Overall death	329 (3.4)	138 (3.4)	1.0	3.7	3.4	0.93 (0.75–1.14)
Palliative care	155 (1.6)	100 (2.5)	0.001	2.0	2.3	1.14 (0.88–1.49)
Transfer to ICU	368 (3.8)	150 (3.7)	0.8	3.6	3.9	1.09 (0.89–1.32)
Transfer to ICU or death	622 (6.5)	248 (6.2)	0.5	6.6	6.3	0.96 (0.82–1.11)
Length-of-stay, d, median (IQR)	5.3 (2.8–9.9)	5.6 (3.0–10.5)	0.001	5.5 (2.8–10.3)	5.6 (2.9–10.5)	–
Secondary outcomes, subspecialty						
Overall death	155 (2.5)	62 (2.6)	0.9	2.7	2.5	0.96 (0.71–1.30)
Palliative care	38 (0.6)	23 (1.0)	0.1	0.8	0.9	1.12 (0.65–1.93)
Transfer to ICU	753 (12.3)	283 (12.0)	0.7	12.8	11.8	0.92 (0.81–1.05)
Transfer to ICU or death	821 (13.5)	322 (13.6)	0.9	14.1	13.3	0.95 (0.84–1.07)
Length-of-stay, d, median (IQR)	5.9 (2.9–12.0)	6.5 (2.9–12.4)	0.3	6.0 (2.9–12.0)	6.4 (2.9–12.2)	–

Note: CI = confidence interval, ICU = intensive care unit, IQR = interquartile range, RR = relative risk.
*Non-palliative, in-hospital death was defined as death that occurred without documented palliative care intervention, identified by the presence of a palliative care diagnosis code or transfer to the inpatient palliative care unit. Palliative care is a composite of deaths with palliative care or transfers to the inpatient palliative care unit.
†Unless indicated otherwise.

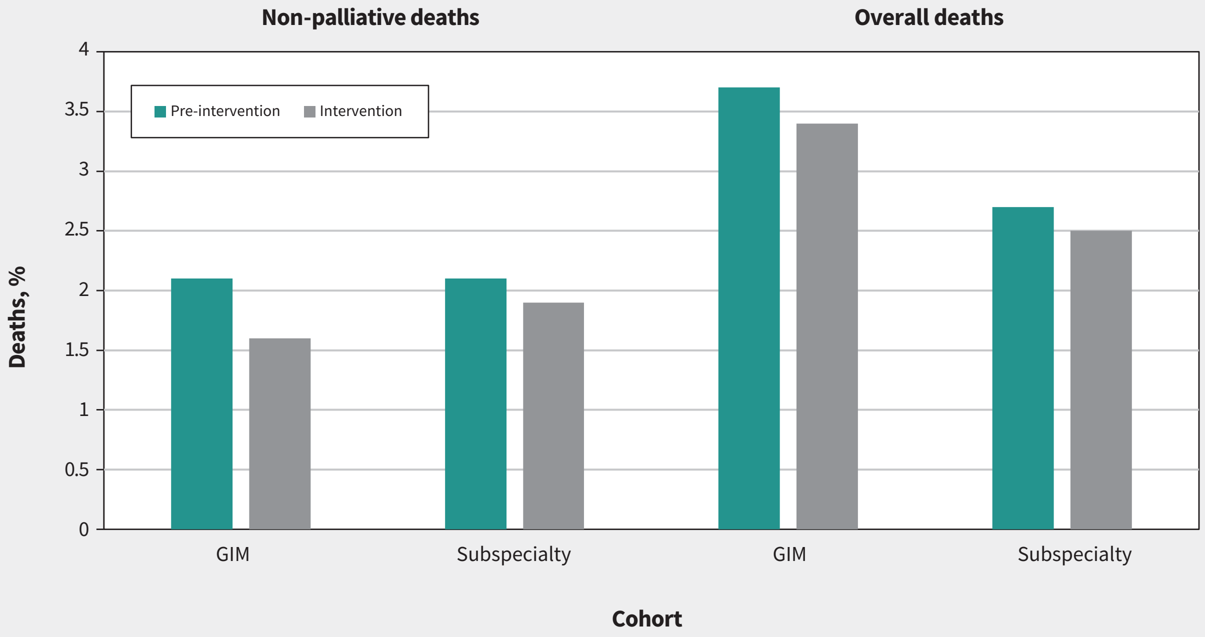

Figure 2: In-hospital deaths during the pre-intervention and intervention periods in the general internal medicine (GIM) and subspecialty cohorts after propensity score–based overlap weighting. Supporting data are presented in Table 2.

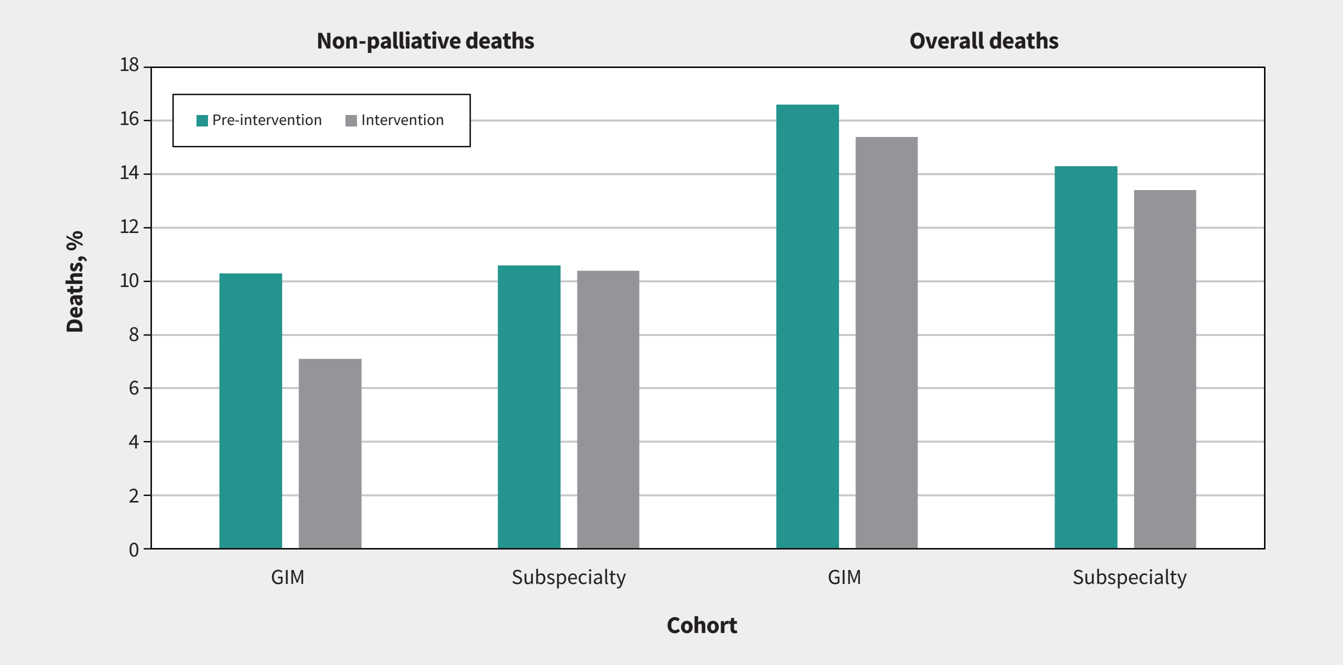

Figure 3: In-hospital deaths during the pre-intervention and intervention periods in the general internal medicine (GIM) and subspecialty high-risk cohorts after propensity score–based overlap weighting. Supporting data are presented in Appendix 1, Table S4, available at www.cmaj.ca/lookup/doi/10.1503/cmaj.240132/tab-related-content.

Table 3: Process of care within 24 hours after the first high-risk alert, after weighting*

	Percentage of patients in GIM†				Percentage of patients in subspecialty†			
Process of care	**Pre-intervention**	**Intervention**	**Difference, %**	***p* value**	**Pre-intervention**	**Contemporaneous**	**Difference, %**	***p* value**
Antibiotics	28.9	49.4	20.5	< 0.001	34.9	47.0	12.1	0.004
Systemic corticosteroid	10.4	17.0	6.6	< 0.001	30.9	38.4	7.5	0.06
Change in code status	12.1	15.0	2.9	0.1	6.4	6.2	–0.2	0.9
Intravenous fluid order	34.5	34.0	–0.5	0.9	11.3	13.9	2.6	0.3
Transfer to step-up unit	3.0	2.4	–0.6	0.6	NA	NA	NA	NA
Radiography	28.5	25.3	–3.2	0.2	1.0	0.7	–0.3	0.7
Computed tomography	12.9	11.8	–1.1	0.6	1.1	0.2	–0.9	0.08
Ultrasonography	9.0	6.8	–2.2	0.2	0.3	< 0.001	–0.3	0.2
Magnetic resonance imaging	4.0	2.1	–1.9	0.06	0.2	< 0.001	–0.2	0.3
No. of vital sign measurements, median (IQR)	3 (2–5)	5 (4–7)	2	< 0.001	4 (3–5)	4 (3–5)	0	0.07

Note: GIM = general internal medicine, IQR = interquartile range, NA = not applicable.
*We included interventions ordered after the high-risk alert. We used the timestamp of when imaging tests were performed to determine if they occurred in the 24 hours after the alert. All interventions in the 24 hours after the first high-risk alert were included, irrespective of whether a patient was transferred to the intensive care unit (interventions could have occurred before or after transfer), since we felt these were all reflective of care that had been delivered and could plausibly have been influenced by a high-risk alert.
†Unless indicated otherwise.

Difference-in-differences analysis

After overlap weighting, the difference in non-palliative deaths in the GIM group was not significantly different than the respective difference in the subspecialty group among all patients (RR reduction 0.79, 95% CI 0.50–1.24) and among high-risk patients (RR reduction 0.66, 95% CI 0.37–1.20). The difference in overall deaths in the GIM group was also not significantly different than the difference in the subspecialty group overall (RR reduction 0.85, 95% CI 0.59–1.22, $p = 0.35$) or in the high-risk group (RR reduction 0.84, 95% CI 0.52–1.35).

Interpretation

The implementation of a multifaceted, machine learning–based early warning system in GIM at an academic centre was associated with lower non-palliative, in-hospital deaths than before implementation. Subspecialty wards, which did not receive the intervention, did not observe a significant change across the same time periods. A difference-in-differences comparison between GIM and subspecialty units demonstrated no statistically significant difference in outcomes, although CIs were wide and did not rule out a clinically meaningful benefit. The outcome of non-palliative death was prespecified and reflected the intended aims of the early warning system, which were to prevent death when possible and improve access to high-quality end-of-life care. Overall, these results suggest that our approach to implementing a machine learning–based early warning system holds promise for improving clinical outcomes but they should be interpreted with caution because of the potential for unmeasured confounding and limited sample size.

Early warning scores are widely used in hospitals. Three recent systematic reviews identified at least 37 studies of machine learning–based early warning systems.[4,6,11] All reviews reported important methodological limitations in the literature, including poor reporting, small sample sizes, insufficient validation, and inconsistent use of model performance and outcome metrics. Few studies have explored the implementation of machine learning–based early warning systems. In 2024, van der Vegt and colleagues[11] systematically reviewed studies of deployed systems and identified key uncertainties in the field, including what type of machine learning model should be used, how predictions should be reported to clinicians, and how early warning systems should integrate with clinical workflow.

Our study offers insights to address these uncertainties. We previously compared various machine learning modelling approaches, including deep learning and simpler machine learning models,[13,14] and found that the simpler MARS models performed similarly to the deep learning neural network models. Here, we reported the clinical effects associated with the real-world deployment of the simpler MARS model, strengthening the general observation that simpler modelling approaches can be effective in the prediction of patient deterioration. In this study, we described in detail how the machine learning model was trained and validated and how machine learning model predictions were incorporated into clinical workflows (Appendix 1). Our system directly engages clinicians through real-time alerts, twice-daily emails to charge nurses to inform nurse–patient assignments, and daily emails to the palliative care team. We developed a care pathway for high-risk patients that focuses on increasing nurse monitoring, enhancing communication between nurses and physicians, and encouraging physicians to reassess patients. Our system does not require remote nurse monitoring as in the Kaiser Permanente intervention,[12] which is important for generalizability, given that many centres would not have the resources to sustain remote nurse monitoring. To reduce alert fatigue, we employed several alert silencing rules.

Little is known about the effect of early warning systems on processes of care.[10,11] In the systematic review by van der Vegt and colleagues,[11] only a single study reported on detailed processes of care and found an increase in oximetry and calls to the primary team and, in contrast to our study, fewer antibiotics after implementation of an early warning system.[20] The only other metrics reported by more than 1 group in the systematic review were ICU transfer rates and time from alert to clinical escalation.[11] We advance this literature by evaluating numerous process measures. We found that the intervention was associated with significantly greater prescribing of antibiotics and corticosteroids, as well as more frequent vital sign monitoring, compared with the pre-intervention period. These findings suggest that the intervention was associated with closer patient monitoring and treatments that could reduce deterioration. More palliative deaths or transfers occurred after the intervention, but this trend was not statistically significant, and fewer deaths occurred overall, although this was also not significant. This suggests that an increase in palliative care alone does not fully explain the observed association with fewer non-palliative deaths. We observed a small absolute increase in ICU transfers among low-risk patients in GIM (around 0.6%) but no increase in deaths, suggesting that resource diversion toward high-risk patients did not negatively affect their care or reduce the likelihood of care escalation.

Limitations

We did not include a randomized control group. The risk of unmeasured confounding is especially important because the first wave of the COVID-19 pandemic affected our hospital 6 months before the intervention. We attempted to account for this by using rich clinical data and propensity score–based weighting, by excluding patients with COVID-19 or influenza, and by using subspecialty medical wards as a contemporaneous control, but this remains an important limitation. Patients admitted to hospital for common non-COVID-19 medical conditions in Ontario and Alberta had similar or greater 30-day mortality during the pandemic (April 2020 to September 2021) than before the pandemic.[36] This suggests that a reduction in deaths during our study period would have been unlikely based on general trends in the province, which is consistent with our observations from subspecialty control units. There was no contamination of patients, most responsible physicians, or interprofessional teams across the GIM and subspecialty units. Control units did not receive early warning alerts or a protocol for high-risk patients. We did not include patients in GIM who were admitted to off-service units (i.e., bedspaced)[37] in the analysis. Our study may be underpowered. We initially planned to compare a 1-year intervention period with a 3-year pre-intervention period, which we expected would yield around

4000 patients in the intervention period and 12 000 control patients, giving 80% power to detect a 1% absolute reduction in deaths from a baseline of around 5% to an anticipated intervention rate of 4%. However, the event rate was lower than expected (in part because of the exclusion of patients with COVID-19 or influenza, although we were not able to distinguish admissions for versus with COVID-19) and we had not planned a priori to compare with subspecialty controls. We chose to extend the study to June 2022 to increase the sample size, but not beyond, as the machine learning model was adjusted at that time because of a change in data pipelines and software tools. The performance of the original model was not a concern, but including predictions from a second model would have complicated our statistical analysis because a single model could not be used to predict patient risk across the different patient cohorts. This was a single-centre study on a GIM unit, limiting generalizability to other specialties and centres. Finally, we were not able to capture all important measures, such as cardiac arrests, code-blue calls, the timing or quality of palliative care, or patient experiences. Future research will explore equity-related considerations related to the intervention and the qualitative experiences of clinical team members.

Conclusion

The implementation of a machine learning–based early warning system was associated with lower non-palliative hospital deaths in GIM after deployment, compared with before system implementation. These findings should be interpreted with caution because of the potential for unmeasured confounding. Our results can inform the approach to implementing machine learning–based early warning systems, which are promising technologies for improving patient outcomes.

References

1. van Galen LS, Struik PW, Driesen BEJM, et al. Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: a root cause analysis of unplanned ICU admissions. *PLoS One* 2016;11:e0161393. doi: 10.1371/journal.pone.0161393.
2. Liu V, Kipnis P, Rizk NW, et al. Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. *J Hosp Med* 2012;7:224-30.
3. Dahn CM, Manasco AT, Breaud AH, et al. A critical analysis of unplanned ICU transfer within 48 hours from ED admission as a quality measure. *Am J Emerg Med* 2016;34:1505-10.
4. 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.
5. Linnen DT, Escobar GJ, Hu X, et al. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. *J Hosp Med* 2019;14:161-9.
6. Muralitharan S, Nelson W, Di S, et al. Machine learning-based early warning systems for clinical deterioration: systematic scoping review. *J Med Internet Res* 2021;23:e25187. doi: 10.2196/25187.
7. Verma AA, Pou-Prom C, McCoy LG, et al. Developing and validating a prediction model for death or critical illness in hospitalized adults, an opportunity for human-computer collaboration. *Crit Care Explor* 2023;5:e0897. doi: 10.1097/CCE.0000000000000897.
8. National Early Warning Score (NEWS) 2. London (UK): Royal College of Physicians; 2022. Available: https://www.rcp.ac.uk/improving-care/resources/national-early-warning-score-news-2/ (accessed 2021 Jan. 21).
9. *National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS — additional implementation guidance*. London (UK): Royal College of Physicians; 2020:1-12.
10. Blythe R, Parsons R, White NM, et al. A scoping review of real-time automated clinical deterioration alerts and evidence of impacts on hospitalised patient outcomes. *BMJ Qual Saf* 2022;31:725-34.
11. van der Vegt AH, Campbell V, Mitchell I, et al. Systematic review and longitud inal analysis of implementing artificial intelligence to predict clinical deterioration in adult hospitals: what is known and what remains uncertain. *J Am Med Inform Assoc* 2024;31:509-24.
12. Escobar GJ, Liu VX, Schuler A, et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020;383:1951-60.
13. Nestor B, McCoy LG, Verma A, et al. Preparing a clinical support model for silent mode in general internal medicine. *Proc Mach Learn Res* 2020;126:950-72.
14. Pou-Prom C, Murray J, Kuzulugil S, et al. From compute to care: lessons learned from deploying an early warning system into clinical practice. *Front Digit Health* 2022;4:932123. doi: 10.3389/fdgth.2022.932123.
15. Verma AA, Murray J, Greiner R, et al. Implementing machine learning in medicine. *CMAJ* 2021;193:E1351-7.
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378. doi: 10.1136/bmj-2023-078378.
17. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022;377:e070904. doi: 10.1136/bmj-2022-070904.
18. Moura CS, Neville A, Liao F, et al. Validity of hospital diagnostic codes to identify SARS-CoV-2 infections in reference to polymerase chain reaction results: a descriptive study. *CMAJ Open* 2023;11:E982-7.
19. Hamilton MA, Calzavara A, Emerson SD, et al. Validating International Classification of Disease 10th Revision algorithms for identifying influenza and respiratory syncytial virus hospitalizations. *PLoS One* 2021;16:e0244746. doi: 10.1371/journal.pone.0244746.
20. Kollef MH, Chen Y, Heard K, et al. A randomized trial of real-time automated clinical deterioration alerts sent to a rapid response team. *J Hosp Med* 2014;9:424-9.
21. Saxena FE, Bierman AS, Glazier RH, et al. Association of early physician follow-up with readmission among patients hospitalized for acute myocardial infarction, congestive heart failure, or chronic obstructive pulmonary disease. *JAMA Netw Open* 2022;5:e2222056. doi: 10.1001/jamanetworkopen.2022.22056.
22. Kostis WJ, Demissie K, Marcella SW, et al.; Myocardial Infarction Data Acquisition System (MIDAS 10) Study Group. Weekend versus weekday admission and mortality from myocardial infarction. *N Engl J Med* 2007;356:1099-109.
23. Quan H, Li B, Couris CM, et al. Updating and validating the charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011;173:676-82.
24. Clinical classifications software refined (CCSR). Rockville (MD): Agency for Healthcare Research and Quality; modified 2024 Apr. 29. Available: https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp (accessed 2020 Dec. 10).
25. Malecki S, Loffler A, Tamming D, et al. Development and external validation of tools for categorizing diagnosis codes in international hospital data. *Int J Med Inform* 2024:189:105508. doi: 10.1016/j.ijmedinf.2024.105508.
26. Matheson F, Moloney G, van Ingen T. *2021 Ontario marginalization index: user guide*. Toronto: St. Michael's Hospital (Unity Health Toronto); 2023. Available: https://www.publichealthontario.ca/-/media/documents/o/2017/on-marg-userguide.pdf (accessed 2023 Oct. 19).
27. Rajaram A, Thomas D, Sallam F, et al. Accuracy of the preferred language field in the electronic health records of two Canadian hospitals. *Appl Clin Inform* 2020;11:644-9.
28. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol* 2019;188:250-7.
29. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc* 2018;113:390-400.
30. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA* 2014;312:2401-2.
31. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. *J Open Source Softw* 2019;4:1686. doi: 10.21105/joss.01686.
32. Yoshida K, Bartel A, Chipman JJ, et al. tableone: Create "Table 1" to describe baseline characteristics with or without propensity score weights. CRAN R; 2022. Available: https://cran.r-project.org/web/packages/tableone/index.html (accessed 2023 May 30).
33. Zhou T, Tong G, Li F, et al. PSweight: an R package for propensity score weighting analysis. *R J* 2022;14:282-300.
34. Lumley T. Analysis of complex survey samples. *J Stat Softw* 2004;9:1-19.
35. Højsgaard S, Halekoh U, Yan J. The R package geepack for generalized estimating equations. *J Stat Softw* 2006;15:1-11.
36. McAlister FA, Chu A, Qiu F, et al.; CORONA Collaboration. Outcomes among patients hospitalized with non-COVID-19 conditions before and during the COVID-19 pandemic in Alberta and Ontario, Canada. *JAMA Netw Open* 2023;6:e2323035. doi: 10.1001/jamanetworkopen.2023.23035.
37. Zannella VE, Jung HY, Fralick M, et al. Bedspacing and clinical outcomes in general internal medicine: a retrospective, multicenter cohort study. *J Hosp Med* 2022;17:3-10.

Competing interests: Amol Verma, Chloe Pou-Prom, Joshua Murray, and Muhammad Mamdani co-invented CHARTwatch, which was acquired by a start-up company, Signal1, and they have the potential to acquire minority interests in the company in the future, as does Sebnem Kuzulugil. Amol Verma received travel support from the Alberta Machine Intelligence Institute and is a part-time employee of Ontario Health, outside of the submitted work.

This article has been peer reviewed.

Affiliations: St. Michael's Hospital (Verma, Colacci, Bell, Ailon, Friedrich, Kuzulugil, Yang, Lee, Pou-Prom, Mamdani), Unity Health Toronto; Department of Medicine (Verma, Colacci, Ailon, Friedrich, Lee, Mamdani), and Institute of Health Policy, Management, and Evaluation (Verma, Stukel, Colacci, Murray, Mamdani), and Department of Laboratory Medicine and Pathobiology (Verma, Mamdani), University of Toronto; ICES Central (Stukel); Leslie Dan Faculty of Pharmacy (Mamdani), University of Toronto, Toronto, Ont.

Contributors: Amol Verma, Chloe Pou-Prom, and Muhammad Mamdani led the design of the study; Amol Verma, Michael Colacci, Shirley Bell, Jonathan Ailon, Jan Friedrich, Joshua Murray, Sebnem Kuzulugil, Zhen Yang, Yuna Lee, Chloe Pou-Prom, and Muhammad Mamdani led the design and implementation of the CHARTwatch intervention. Chloe Pou-Prom led data collection. Therese Stukel oversaw statistical analysis and Chloe Pou Prom conducted data analysis. All of the authors contributed to data interpretation. Amol Verma drafted the manuscript. All of the authors revised it critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work. Chloe Pou-Prom and Muhammad Mamdani were equal contributors and co–senior authors.

Funding: Amol Verma is supported by the Temerty Professorship of AI Research and Education in Medicine at the University of Toronto. This project was supported in-part by the Vector Institute Pathfinder Project and the AMS Healthcare Compassion and AI Fellowship. Funders played no role in the research.

Data sharing: The data from this study can be accessed upon reasonable request to the corresponding author, and will be provided in line with privacy and ethics requirements. A portion of the data that were used in this study are available in the Health Data Nexus repository (https://healthdatanexus.ai/content/smh-gim/1.0.1/).

Accepted: Aug. 1, 2024

Correspondence to: Amol Verma, amol.verma@mail.utoronto.ca