JAMA Oncology | Original Investigation

Validation of a Machine Learning Algorithm to Predict 180-Day Mortality for Outpatients With Cancer

Christopher R. Manz, MD; Jinbo Chen, PhD; Manqing Liu, MHS; Corey Chivers, PhD; Susan Harkness Regli, PhD; Jennifer Braun, MHA; Michael Draugelis, MS; C. William Hanson, MD; Lawrence N. Shulman, MD; Lynn M. Schuchter, MD; Nina O'Connor, MD; Justin E. Bekelman, MD; Mitesh S. Patel, MD, MBA; Ravi B. Parikh, MD, MPP

IMPORTANCE Machine learning (ML) algorithms can identify patients with cancer at risk of short-term mortality to inform treatment and advance care planning. However, no ML mortality risk prediction algorithm has been prospectively validated in oncology or compared with routinely used prognostic indices.

OBJECTIVE To validate an electronic health record–embedded ML algorithm that generated real-time predictions of 180-day mortality risk in a general oncology cohort.

DESIGN, SETTING, AND PARTICIPANTS This prognostic study comprised a prospective cohort of patients with outpatient oncology encounters between March 1, 2019, and April 30, 2019. An ML algorithm, trained on retrospective data from a subset of practices, predicted 180-day mortality risk between 4 and 8 days before a patient's encounter. Patient encounters took place in 18 medical or gynecologic oncology practices, including 1 tertiary practice and 17 general oncology practices, within a large US academic health care system. Patients aged 18 years or older with outpatient oncology or hematology and oncology encounters were included in the analysis. Patients were excluded if their appointment was scheduled after weekly predictions were generated and if they were only evaluated in benign hematology, palliative care, or rehabilitation practices.

EXPOSURES Gradient-boosting ML binary classifier.

MAIN OUTCOMES AND MEASURES The primary outcome was the patients' 180-day mortality from the index encounter. The primary performance metric was the area under the receiver operating characteristic curve (AUC).

RESULTS Among 24 582 patients, 1022 (4.2%) died within 180 days of their index encounter. Their median (interquartile range) age was 64.6 (53.6-73.2) years, 15 319 (62.3%) were women, 18 015 (76.0%) were White, and 10 658 (43.4%) were seen in the tertiary practice. The AUC was 0.89 (95% CI, 0.88-0.90) for the full cohort. The AUC varied across disease-specific groups within the tertiary practice (AUC ranging from 0.74 to 0.96) but was similar between the tertiary and general oncology practices. At a prespecified 40% mortality risk threshold used to differentiate high- vs low-risk patients, observed 180-day mortality was 45.2% (95% CI, 41.3%-49.1%) in the high-risk group vs 3.1% (95% CI, 2.9%-3.3%) in the low-risk group. Integrating the algorithm into the Eastern Cooperative Oncology Group and Elixhauser comorbidity index–based classifiers resulted in favorable reclassification (net reclassification index, 0.09 [95% CI, 0.04-0.14] and 0.23 [95% CI, 0.20-0.27], respectively).

CONCLUSIONS AND RELEVANCE In this prognostic study, an ML algorithm was feasibly integrated into the electronic health record to generate real-time, accurate predictions of short-term mortality for patients with cancer and outperformed routinely used prognostic indices. This algorithm may be used to inform behavioral interventions and prompt earlier conversations about goals of care and end-of-life preferences among patients with cancer.

Invited Commentary
page 1731

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Ravi B. Parikh, MD, MPP, Department of Medicine, University of Pennsylvania Perelman School of Medicine, 423 Guardian Dr, Blockley 1102, Philadelphia, PA 19104 (ravi.parikh@pennmedicine.upenn.edu).

JAMA Oncol. 2020;6(11):1723-1730. doi:10.1001/jamaoncol.2020.4331
Published online September 24, 2020.

Patients with cancer frequently receive care near the end of life that is not concordant with their wishes and worsens their quality of life.[1-6] In an effort to curb unwarranted end-of-life care, national quality metrics in oncology increasingly focus on treatment intensity and hospice use near the end of life.[7-10] Achieving such metrics relies on clinicians accurately predicting a patient's risk of mortality. However, oncologists are often unable to identify patients' short-term mortality risk, with a bias toward overestimating life expectancy.[11-13] Underestimating mortality risk is associated with aggressive end-of-life care among patients with cancer.[14,15] Better assessment of short-term mortality risk may inform advance care planning and increase goal-concordant cancer care.[16-21]

By using vast longitudinal data from the electronic health record (EHR), machine learning (ML) algorithms may improve mortality risk prediction compared with population-level and randomized controlled trial estimates.[22-24] However, most EHR-based ML algorithms to predict mortality have only been validated on retrospective data from patients receiving older treatment strategies. This raises questions about the generalizability of ML mortality risk prediction algorithms in real-world practice. To ensure accuracy and clinical usefulness, ML mortality risk prediction algorithms must be prospectively validated in large, unselected cohorts of patients with cancer and compared with existing prognostic indices.

In this prognostic cohort study, we describe the prospective performance of a previously trained ML algorithm based on structured EHR data to predict 180-day mortality in outpatients seen in oncology practices within a large academic system.[25] We hypothesized that this algorithm would have good discrimination and calibration and outperform routinely used prognostic indices in oncology.

Key Points

Question Can a machine learning algorithm prospectively identify patients with cancer at risk of 180-day mortality?

Findings In this prognostic cohort study of 24 582 patients seen in oncology practices within a large health care system, a machine learning algorithm integrated into the electronic health record accurately identified the risk of 180-day mortality with good discrimination and positive predictive value of 45.2%. When added to performance status– and comorbidity-based classifiers, the algorithm favorably reclassified patients.

Meaning An integrated machine learning algorithm demonstrated good prospective performance compared with traditional prognostic classifiers and may inform clinician and patient decision-making in oncology.

Methods

Setting

This study was performed at medical or gynecologic oncology practices in Pennsylvania and New Jersey within the University of Pennsylvania Health System (UPHS). Eligible practices included a large tertiary practice, in which clinicians subspecialize in 1 cancer type (eg, lymphoma), and 17 general oncology practices, in which clinicians usually treat multiple cancer types. The UPHS institutional review board approved this study with a waiver of informed consent, classifying the study as quality improvement. This study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline for prediction algorithm validation.[26]

Study Population

Patients were eligible for this study if they were aged 18 years or older and had an encounter between March 1, 2019, and April 30, 2019 (the baseline period), that was listed in Clarity, an Epic reporting database that contains structured data elements of individual EHRs for patients within the UPHS.[27] As this was a pragmatic study, the sample size was determined by the number of patient encounters during the baseline period and not by a priori power calculations. For patients with multiple encounters during the baseline period, the last encounter in the baseline period was defined as the index encounter. Patients were followed up until death. Patients who were living at 180 days from the index encounter were censored in analyses. Of 25 588 patients who had any eligible encounter during the baseline period, 24 582 (96.1%) were eligible for this study. Excluded patients included 867 (3.4%) whose appointment was scheduled after weekly predictions were generated and 139 (0.5%) who were only seen in benign hematology, palliative care, or rehabilitation practices.

Study Design

This study prospectively evaluated a gradient-boosted ML algorithm used to predict 180-day mortality among outpatients with cancer. We previously trained this algorithm, which uses 559 structured EHR features as inputs, using historic data from 2016.[25] Of note, certain variables, including Eastern Cooperative Oncology Group (ECOG) performance status, stage, and genetic variants, were not included as features in the algorithm because they were not reliably coded in structured format at the time of algorithm training. Algorithm specifications, features, handling of missing data, cross-validation, and source code are described in this previous publication. The algorithm was integrated into the EHR at UPHS in early 2019. Of the 18 eligible practices in this study, 7 medical oncology practices were part of the original algorithm training data set, and 11 medical and gynecologic oncology practices were not part of the original algorithm training data set. The algorithm was not recalibrated after training; thus, all features and weights were locked before implementation into the UPHS EHR as part of this prospective cohort study. Importantly, no patient data used in this study were used in the original algorithm training.

In this study, patients who had an encounter in 1 of 18 practices during the baseline period were followed up for 180 days after the index encounter. We selected a 60-day baseline period during which the algorithm ran silently for all encounters and clinicians were not exposed to algorithm predictions; thus, algorithm output could not immediately affect care

patterns (eg, early stoppage of chemotherapy, early referral to hospice) that could have influenced the outcome of 180-day mortality. Beginning shortly after this 60-day period, select clinicians began being exposed to predictions as part of a randomized controlled trial.[28] During the baseline period, a database of all required structured EHR data was updated nightly. The algorithm engine ran automatically once a week on Thursdays at 7:00 AM and used EHR data updated on the previous night to generate risk predictions of 180-day mortality for each patient whose encounter was scheduled for the coming Monday through Friday during the baseline period. Patient encounters that were scheduled after the prior Thursday at 7:00 AM did not have an associated prediction and were not included in the cohort.

Outcome

The primary outcome of this study was 180-day mortality, defined as any death within the 180 days after the index encounter. Cause of death was not available for this analysis. Date of death was derived from the first date of death recorded in either the EHR or the Social Security Administration Death Master File and matched to patients within the UPHS by social security number and date of birth.[29] All dates of death were ascertained on January 20, 2020.

Statistical Analysis

We used descriptive statistics to compare characteristics of the cohort, stratified by site of care (tertiary academic vs general oncology and practices that were included in the algorithm training data set vs practices that were not included in the algorithm training data set). To assess algorithm discrimination, we calculated the area under the receiver operating characteristic curve (AUC) as our primary performance metric. To account for class imbalance due to low overall rates of 180-day mortality, we also calculated the area under the precision-recall curve (AUPRC).[30] Algorithm calibration was assessed by plotting the predicted vs observed rate of 180-day mortality and by calculating the scaled Brier score, which is a measure of calibration ranging from 0 (poor calibration) to 1 (perfect calibration).[31] To assess other performance metrics of operational importance, we calculated positive predictive value (PPV), negative predictive value (NPV), sensitivity, and alert rate (the percentage of all encounters deemed as high risk) at various risk thresholds. All threshold-dependent performance metrics in the primary analysis are presented at a prespecified 40% risk of 180-day mortality unless otherwise specified, because this was the threshold used to report the performance of the algorithm after training and was determined after consultation with select oncology clinicians.[25] To assess for heterogeneity of algorithm performance, we calculated performance metrics across important prespecified subgroups: disease site, practice type (tertiary vs general), self-reported race, sex, insurance type, and stage (American Joint Committee on Cancer 8th edition I-III vs IV).

There is no criterion standard prognostic index for patients with cancer. To compare the ML algorithm with other commonly used prognostic references in general oncology, we used logistic regression to create 2 univariable classifiers based on ECOG (ECOG ≥2 vs <2) and number of Elixhauser comorbidities (Elixhauser ≥3 vs <3).[32,33] We calculated the aforementioned performance metrics for these univariable classifiers based on ECOG values that were coded less than or equal to 30 days before the index encounter and Elixhauser comorbidities coded before the index encounter, respectively. Patients without coded ECOG values were not used to calculate the performance of the ECOG classifier. The Elixhauser comorbidity index was available for all patients. To compare the ML algorithm against these 2 indices, we calculated 3 metrics: (1) the difference in AUC (ΔAUC), (2) the difference in PPV (ΔPPV), and (3) the additive net reclassification index (NRI).[34] The NRI measures the improvement in predictive performance gained by adding the ML algorithm to existing prognostic indices; positive values of the NRI indicate overall correct reclassification of true positives and true negatives, whereas negative values indicate incorrect reclassification. To calculate the NRI, we first developed a logistic regression algorithm with mortality (dead vs alive) as the outcome and baseline classifier (ECOG or Elixhauser) as the only covariate. We then developed another logistic regression algorithm (enhanced classifier) that added the ML classifier (prediction score from ML algorithm fourth quartile vs first to third quartile) as a covariate to the baseline classifiers. We used the fourth quartile of the prediction score from the ML algorithm as the threshold to classify patients into high- and low-risk groups and calculated the NRI based on the number of patients who were reclassified.

All analyses were conducted using the Sklearn, version 0.21.3 package in Python (Python Software Foundation); all hypothesis testing used 2-sided $P < .05$ to determine statistical significance. Statistical analysis was performed between January 1, 2020, and March 15, 2020.

Sensitivity Analysis

For patients with multiple encounters during the baseline period, we repeated analyses using the first and a random encounter as the index encounter. To assess threshold-dependent performance metrics, we used 10%, 20%, and 30% risk of 180-day mortality as the risk threshold and recalculated performance metrics.

Results

The cohort consisted of 24 582 patients, of whom 1022 (4.2%) died within 180 days of their index encounter. In the full cohort, the median (interquartile range) age was 64.6 (53.6-73.2) years, 15 319 (62.3%) were women, 18 015 (76.0%) were White, and 10 658 (43.4%) were seen in the tertiary practice. A total of 15 059 patients (61.3%) were married, and 6754 (27.5%) had an Elixhauser count of greater than or equal to 3. There were large differences between the patients who were alive and those who were deceased within 180 days of the index encounter (**Table 1**). Compared with living patients, deceased patients were more likely to have documented stage IV cancer (136 [59.1%] vs 835 [22.2%]), ECOG greater than or equal to 2 (261 [56.5%] vs 702 [12.2%]), and Elixhauser count greater

Table 1. Study Population Characteristics

Characteristic	No. (%)[a]		
	Total (n = 24 582)	Alive (n = 23 560)	Dead (n = 1022)
Age, median (IQR), y	64.6 (53.6-73.2)	64.4 (53.3-73.0)	69.7 (60.5-78.0)
Women	15 319 (62.3)	14 803 (62.8)	516 (50.5)
Race/ethnicity			
Non-Hispanic White	18 015 (76.0)	17 263 (76.0)	752 (75.8)
Black	3951 (16.7)	3769 (16.6)	182 (18.3)
Other	1734 (7.3)	1676 (7.4)	58 (5.8)
Marital status			
Married	15 059 (61.3)	14 451 (61.4)	608 (59.6)
Unmarried	9493 (38.7)	9081 (38.6)	412 (40.4)
Insurance			
Medicare + managed Medicare	12 244 (50.8)	2409 (10.4)	77 (7.8)
Medicaid + managed Medicaid	1247 (5.2)	7917 (34.2)	201 (20.4)
Managed care	8118 (33.7)	1178 (5.1)	69 (7.0)
Commercial insurance	2486 (10.3)	11 608 (50.2)	636 (64.7)
ECOG performance status			
0	2822 (45.3)	2785 (48.3)	37 (8.0)
1	2440 (39.2)	2276 (39.5)	164 (35.5)
≥2	963 (15.5)	702 (12.2)	261 (56.5)
Elixhauser comorbidity score			
0-1	12 866 (52.3)	12 695 (53.9)	171 (16.7)
2	4962 (20.2)	4764 (20.2)	198 (19.4)
≥3	6754 (27.5)	6101 (25.9)	653 (63.9)
Cancer stage			
I-III	3028 (75.7)	2934 (77.8)	94 (40.9)
IV	971 (24.3)	835 (22.2)	136 (59.1)
Practice site			
Tertiary academic	10 658 (43.4)	10 136 (43.0)	522 (51.1)
Breast	1862 (11.0)	1811 (11.3)	51 (6.3)
Gastrointestinal tract	1193 (7.1)	1074 (6.7)	119 (14.7)
Thoracic	1060 (6.3)	961 (6.0)	99 (12.3)
Genitourinary	1048 (6.2)	1009 (6.3)	39 (4.8)
Gynecologic	1049 (6.2)	1029 (6.4)	20 (2.5)
Leukemia	1062 (6.3)	993 (6.2)	69 (8.6)
Lymphoma	1259 (7.5)	1214 (7.6)	45 (5.6)
Melanoma	403 (2.4)	381 (2.4)	22 (2.7)
Myeloma	1179 (7.0)	1138 (7.1)	41 (5.1)
Neuro-oncology	355 (2.1)	333 (2.1)	22 (2.7)
General oncology	13 924 (56.6)	13 424 (57.0)	500 (48.9)

Abbreviations: ECOG, Eastern Cooperative Oncology Group; IQR, interquartile range.

[a] Values are presented as No. (%) unless otherwise specified.

than or equal to 3 (653 [63.9%] vs 6101 [25.9%]). Full characteristics of patients, stratified by practice site, inclusion in original algorithm training data, and ECOG documentation status, are available in eTables 1, 2, and 3 in the Supplement, respectively.

Algorithm Performance

Algorithm performance metrics in the full cohort and stratified by practice site are described in **Table 2**. In the full cohort, the AUC was 0.89 (95% CI, 0.88-0.90), the AUPRC was 0.34, and the scaled Brier score was 0.29. Algorithm performance did not differ between patients seen at the tertiary academic vs general oncology practices. Among practice sites that were not part of the original algorithm training, the AUC was 0.86. Performance characteristics varied across disease-specific groups within the tertiary practice, ranging from an AUC of 0.74 (neuro-oncology) to 0.96 (breast oncology). Algorithm performance was slightly better for women compared with men (AUC, 0.91 vs 0.86; AUPRC, 0.37 vs 0.31; PPV, 0.47 vs 0.43; sensitivity, 0.30 vs 0.25) and for patients with stage IV compared with stage I to III cancer (AUPRC, 0.47 vs 0.24; PPV, 0.49 vs 0.31; sensitivity, 0.39 vs 0.25) (eTable 4 in the Supplement). There were no significant differences in performance across race/ethnicity and insurance status.

At a prespecified mortality risk threshold of 40%, 2.5% of encounters were flagged as high risk. At this threshold, the PPV

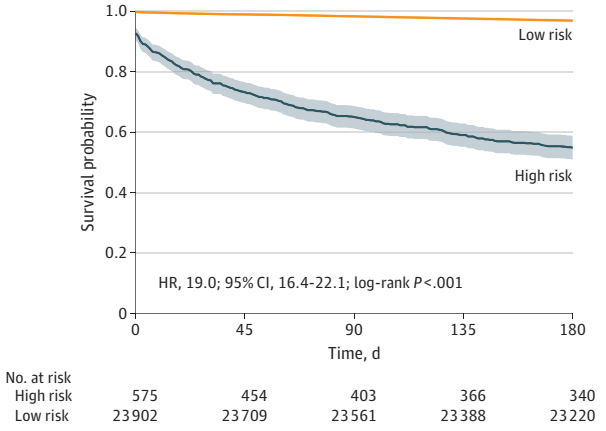
Table 2. Algorithm Performance

Variable	AUC (95% CI)	AUPRC	Scaled Brier score	PPV	NPV	Sensitivity	Specificity	No.	Rate	
									Mortality	Alert
Overall	0.89 (0.88-0.90)	0.34	0.29	0.45	0.97	0.27	0.99	24 582	0.04	0.025
Practice site[a]										
Included in algorithm training										
Yes	0.89 (0.88-0.90)	0.37	0.29	0.47	0.97	0.29	0.98	18 927	0.05	0.03
No	0.86 (0.83-0.88)	0.22	0.26	0.36	0.98	0.17	0.99	5655	0.03	0.01
Practice site										
Tertiary academic	0.89 (0.87-0.90)	0.37	0.31	0.45	0.97	0.31	0.98	10 658	0.05	0.03
Breast	0.96 (0.94-0.98)	0.49	0.38	0.57	0.99	0.41	0.99	1862	0.03	0.02
Gastrointestinal	0.85 (0.81-0.88)	0.38	0.32	0.39	0.92	0.29	0.95	1193	0.10	0.08
Thoracic	0.82 (0.77-0.86)	0.38	0.32	0.42	0.93	0.34	0.95	1060	0.09	0.08
Genitourinary	0.88 (0.80-0.94)	0.33	0.43	0.39	0.98	0.36	0.98	1048	0.04	0.03
Gynecologic	0.91 (0.87-0.95)	0.24	0.25	0.50	0.99	0.25	1.00	1049	0.02	0.01
Leukemia	0.85 (0.79-0.90)	0.39	0.27	0.50	0.96	0.33	0.98	1062	0.07	0.04
Lymphoma	0.91 (0.85-0.95)	0.29	0.14	0.50	0.97	0.16	1.00	1259	0.04	0.01
Melanoma	0.90 (0.84-0.95)	0.35	0.20	0.25	0.95	0.09	0.98	403	0.06	0.02
Myeloma	0.91 (0.85-0.96)	0.44	0.26	0.62	0.98	0.32	0.99	1179	0.04	0.02
Neuro-oncology	0.74 (0.64-0.83)	0.14	0.21	0.11	0.94	0.05	0.98	355	0.06	0.03
General oncology	0.89 (0.87-0.90)	0.31	0.27	0.45	0.97	0.24	0.99	13 924	0.04	0.02

Abbreviations: AUC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; PPV, positive predictive value; NPV, negative predictive value.

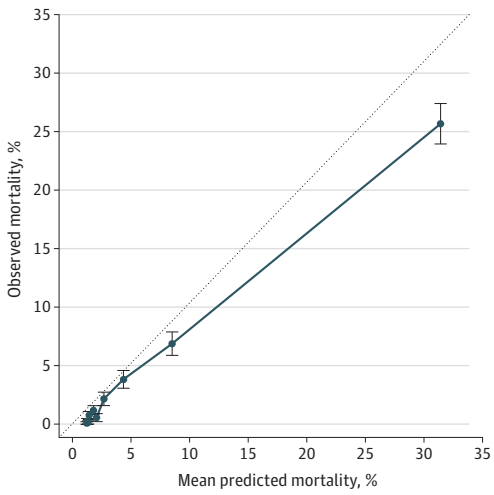
[a] These data only refer to the site's inclusion in the original algorithm training (based on 2016 data). No individual patient data used in the original algorithm training were reused in the prospective validation.

Figure 1. Overall 180-Day Mortality of Patients Considered High vs Low Risk as Identified by a Machine Learning Algorithm



High-risk patients defined as having greater than 40% risk of 180-day mortality.

Figure 2. Algorithm Calibration



Points are binned deciles of predicted mortality. Mean predicted mortality refers to predicted risk of 180-day mortality from the machine learning algorithm. Observed mortality refers to the percentage of patients who died within 180 days of the index encounter. Bars refer to 95% confidence intervals of the prediction. The dotted line refers to a reference model where mean predicted mortality equals observed mortality; a perfectly calibrated algorithm would fall along the dotted line. Points falling below the dotted line overestimate risk of mortality, whereas points falling above the dotted line underestimate risk of mortality.

was 45.2%, NPV was 96.9%, and sensitivity was 27.4%. Observed 180-day mortality was 45.2% (95% CI, 41.3%-49.1%) in the high-risk group vs 3.1% (95% CI, 2.9%-3.3%) in the low-risk group (**Figure 1**); high-risk patients had higher 180-day mortality across varying thresholds of predicted risk (eFigure 1 in the Supplement). The algorithm was well calibrated for patients with mortality risk less than or equal to 10% and overestimated mortality risk for patients with mortality risk greater than 10% (**Figure 2**). Observed mortality was 100-fold higher among patients in the highest vs lowest quartiles of predicted risk (13.2% vs 0.13%) (eFigure 2 in the Supplement).

Algorithm Comparison

Of patients who had a coded ECOG (6225 [25.3%]), 963 (15.5%) had an ECOG greater than or equal to 2. Compared with the

Table 3. Comparison of ML Algorithm Against Standard Prognostic Indices in Oncology

Variable	Prognostic index	
	ECOG (n = 6225)	Elixhauser (n = 24 582)
Baseline classifier AUC (95% CI)	0.72 (0.70-0.75)	0.69 (0.68-0.71)
ΔAUC (95% CI) between ML algorithm and baseline classifier	0.17 (0.14-0.19)	0.20 (0.18-0.21)
Baseline classifier PPV, No.	0.27	0.10
ΔPPV between ML algorithm and baseline classifier, No.	0.18	0.36
NRI (95% CI) from enhanced vs baseline classifier	0.09 (0.04-0.14)	0.23 (0.20-0.27)

Abbreviations: AUC, area under the receiver operating characteristic curve; ECOG, Eastern Cooperative Oncology Group; ML, machine learning; NRI, net reclassification index; PPV, positive predictive value; ΔAUC, difference in AUC; ΔPPV, difference in PPV.

baseline ECOG-only classifier, the enhanced classifier integrating the ECOG and ML classifiers had a significantly higher AUC (ΔAUC, 0.17; 95% CI, 0.14-0.19) and higher PPV (ΔPPV, 0.18). The enhanced classifier resulted in favorable reclassification compared with the baseline classifier (NRI, 0.09; 95% CI, 0.04-0.14) (**Table 3**).

Compared with the baseline Elixhauser-only classifier, the enhanced classifier integrating the Elixhauser and ML classifiers had a significantly higher AUC (ΔAUC, 0.20; 95% CI, 0.18-0.21) and higher PPV (ΔPPV, 0.36). The enhanced classifier resulted in favorable reclassification compared with the baseline classifier (NRI, 0.23; 95% CI, 0.20-0.27) (Table 3).

Subgroup and Sensitivity Analyses

Overall algorithm performance characteristics were similar when, for patients with multiple encounters in the database, either the first or a random encounter was used as the index encounter (eTable 5 in the Supplement). Threshold-dependent performance characteristics varied when alternative mortality risk thresholds were applied (eTable 6 in the Supplement).

Discussion

In this prognostic cohort study, our results suggest that an ML algorithm can be feasibly integrated into the EHR to generate real-time, accurate predictions of short-term mortality risk for patients with cancer that outperform traditional prognostic indices. This study represents, to our knowledge, one of the first prospective applications of an ML mortality risk prediction algorithm in clinical oncology and provides 3 important insights into the potential clinical applicability of real-time prognostic algorithms in clinical practice.

First, this study is a true prospective validation of a previously described ML algorithm. The ML algorithm was trained using patient data from 2016. In contrast, this prospective cohort was enrolled in 2019. Thus, no patient data in this prospective validation was used as part of algorithm training. Despite many therapeutic advances since 2016, including the widespread use of immunotherapy and biomarker-based therapies, our algorithm demonstrated good performance on several clinically relevant metrics even when applied to a prospective cohort 3 years later.[35] Additionally, the AUC, AUPRC, and PPV from our prospective validation were equivalent to or better than what has been reported in retrospective validations of ML mortality risk prediction algorithms in oncology.[23,25,36] This demonstrates that an ML mortality risk prediction algorithm trained on retrospective data may still have good performance and clinical utility when applied in a recent cohort of patients with cancer. Importantly, the algorithm still had a clinically applicable AUC (0.86) for practices that were not included in the algorithm training.

Second, the ML algorithm demonstrated better AUC and PPV compared with the ECOG and Elixhauser classifiers, and enhanced classifiers that integrated the ML algorithm into existing prognostic indices resulted in better reclassification of patients. Published ML prognostic algorithms often either do not have a comparator or have been compared with regression-based algorithms that are not used in clinical practice.[25,37] In contrast, ECOG and Elixhauser comorbidity scores are validated prognostic indices that are commonly used in decision-making around treatment selection and clinical trial enrollment.[38] Although Elixhauser comorbidities are included in the ML algorithm inputs, ECOG is not included in the algorithm because it was not available as a structured data element in the EHR at the time of algorithm training.[25] As ECOG and other variables of prognostic significance (eg, stage, line of treatment, genetic variants) are increasingly coded in the EHR, the performance of this algorithm could improve.

Third, the differences in prospective performance across cancer types and stages give important information regarding for which disease settings an automated ML prognostic algorithm may be most useful in general oncology. For malignancies with poorer performance (eg, melanoma, neuro-oncology), several factors that were not included in our algorithm may have greater prognostic importance; these include performance and cognitive status, molecular and genetic biomarker expression, and imaging results. When considering the implementation of a tool such as this, clinicians must consider the need to account for prognostic factors that are unique to specific cancers. Relatedly, clinicians may use this point-of-care tool differently with different threshold preferences for patients with stage I to III cancer—who are generally treated with curative intent—compared with patients with stage IV cancer who are generally treated with palliative intent. Finally, despite recent concerns over algorithm bias against underrepresented subgroups,[39,40] there were no major differences in performance across site of practice, race/ethnicity, and insurance status.

Although ML algorithms may improve mortality risk prediction, it is important to demonstrate that improved accuracy can translate to better clinician and patient decision-making.[41] ML mortality risk prediction tools may help oncology clinicians better identify patients who are at high risk of short-term mortality, engage in early goals of care discussions, and meet quality metrics related to end-of-life care. Indeed, in a subsequent randomized clinical trial, we applied this

algorithm in clinical practice to identify oncology patients with the highest predicted mortality and used behavioral nudges to encourage clinicians to have end-of-life conversations with those patients, resulting in significant increases in such conversations.[28,42] Notably, for a cancer with a high baseline rate of short-term mortality (eg, pancreatic cancer), the mortality risk threshold for intervention may be higher to avoid a high alert rate; the opposite may be true for a cancer with a low rate of short-term mortality (eg, prostate cancer, myeloma). Our algorithm retained good performance across several mortality risk thresholds, indicating that such thresholds could be customized for clinical use (eTable 6 in the Supplement).

Limitations

There are several limitations to this study. First, using a risk threshold of 40%, the sensitivity of the algorithm was only 27%. This may reflect the inability to include important features, such as stage and ECOG values, in the original algorithm training, in addition to the high mortality risk threshold used. Indeed, when using a lower but still clinically meaningful risk threshold of 10%, sensitivity improved to nearly 70% (eTable 6 in the Supplement). Third, in the comparison of the ML algorithm against the ECOG classifier, we were limited to only using coded ECOG values. Patients whose ECOG was coded were systematically different than patients whose ECOG was not coded, as shown in eTable 3 in the Supplement. However, the purpose of this analysis was not to develop a new classifier that includes ECOG but rather to assess whether the ML algorithm improves classification above and beyond the ECOG for those with a coded ECOG. Indeed, the positive NRI indicates that even within the subgroup of those with coded ECOG, the ML algorithm led to favorable reclassification of patients compared with the ECOG classifier alone. Using imputation strategies would not have been appropriate given the high degree of missingness of ECOG.

Conclusions

The results of this prognostic study suggest that an ML algorithm to predict short-term mortality has good prospective validity and compares favorably with existing prognostic indices. Such an automated tool may complement clinician intuition and lead to improved targeting of supportive care interventions for high-risk patients with cancer.

ARTICLE INFORMATION

Accepted for Publication: July 6, 2020.

Published Online: September 24, 2020. doi:10.1001/jamaoncol.2020.4331

Author Affiliations: Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia (Manz, Shulman, Schuchter, O'Connor, Bekelman, Patel, Parikh); Penn Center for Cancer Care Innovation, University of Pennsylvania, Philadelphia (Manz, Shulman, Schuchter, Bekelman, Patel, Parikh); Abramson Cancer Center, University of Pennsylvania, Philadelphia (Manz, Braun, Shulman, Schuchter, Bekelman, Parikh); Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia (Manz, Liu, Bekelman, Patel, Parikh); Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia (Chen); Penn Medicine, University of Pennsylvania, Philadelphia (Chivers, Regli, Draugelis, Hanson); Department of Medical Ethics and Health Policy, University of Pennsylvania Perelman School of Medicine, Philadelphia (Bekelman, Patel, Parikh); Corporal Michael J. Crescenz VA Medical Center, Philadelphia, Pennsylvania (Patel, Parikh).

Author Contributions: Dr Parikh and Ms Lin had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Manz, Chen, Chivers, Regli, Braun, Draugelis, Hanson, Shulman, Schuchter, O'Connor, Parikh.
Acquisition, analysis, or interpretation of data: Manz, Chen, Liu, Chivers, Braun, Schuchter, O'Connor, Bekelman, Patel, Parikh.
Drafting of the manuscript: Manz, Chen, Liu, Parikh.
Critical revision of the manuscript for important intellectual content: Manz, Chen, Chivers, Regli, Braun, Draugelis, Hanson, Shulman, Schuchter, O'Connor, Bekelman, Patel, Parikh.
Statistical analysis: Manz, Chen, Liu, Chivers, Parikh.
Obtained funding: Schuchter, Parikh.
Administrative, technical, or material support: Manz, Chivers, Regli, Braun, Draugelis, Shulman, O'Connor, Parikh.
Supervision: Manz, Regli, Draugelis, Hanson, Shulman, Schuchter, O'Connor, Patel, Parikh.

Conflict of Interest Disclosures: Dr Bekelman reported receiving grants from Pfizer, UnitedHealth Group, North Carolina Blue Cross Blue Shield, and Embedded Healthcare; personal fees from the Centers for Medicare and Medicaid Services, Optum, CVS Health, National Comprehensive Cancer Network, and UnitedHealthcare outside the submitted work. Dr Patel reported receiving personal fees from Catalyst Health LLC, HealthMine Services, and Holistic Industries outside the submitted work. Dr Parikh reported receiving personal fees from GNS Healthcare and Cancer Study Group, grants and nonfinancial support from Conquer Cancer Foundation, and grants from MUSC Transdisciplinary Collaborative Center in Precision Medicine and Minority Men's Health, VA Center for Health Equity Research and Promotion, National Palliative Care Research Center, Embedded Healthcare, and University of Pennsylvania Institute for Translational Medicine and Therapeutics outside the submitted work. No other disclosures were reported.

Funding/Support: This study was supported by a grant from the Penn Center for Precision Medicine Accelerator Fund (Drs Manz and Parikh) and grant T32 GM075766-14 from the National Institute of General Medical Sciences (Dr Manz).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Meeting Presentation: This study was presented as a poster discussion at the American Society of Clinical Oncology 2020 Annual Meeting; May 29, 2020. This meeting was held online owing to the COVID-19 pandemic.

REFERENCES

1. Earle CC, Neville BA, Landrum MB, Ayanian JZ, Block SD, Weeks JC. Trends in the aggressiveness of cancer care near the end of life. *J Clin Oncol*. 2004; 22(2):315-321. doi:10.1200/JCO.2004.08.136

2. Chastek B, Harley C, Kallich J, Newcomer L, Paoli CJ, Teitelbaum AH. Health care costs for patients with cancer at the end of life. *J Oncol Pract*. 2012;8 (6):75s-80s. doi:10.1200/JOP.2011.000469

3. Wright AA, Mack JW, Kritek PA, et al. Influence of patients' preferences and treatment site on cancer patients' end-of-life care. *Cancer*. 2010;116(19): 4656-4663. doi:10.1002/cncr.25217

4. Detering KM, Hancock AD, Reade MC, Silvester W. The impact of advance care planning on end of life care in elderly patients: randomised controlled trial. *BMJ*. 2010;340:c1345. doi:10.1136/bmj.c1345

5. Wright AA, Zhang B, Ray A, et al. Associations between end-of-life discussions, patient mental health, medical care near death, and caregiver bereavement adjustment. *JAMA*. 2008;300(14): 1665-1673. doi:10.1001/jama.300.14.1665

6. Earle CC, Park ER, Lai B, Weeks JC, Ayanian JZ, Block S. Identifying potential indicators of the quality of end-of-life cancer care from administrative data. *J Clin Oncol*. 2003;21(6):1133-1138. doi:10.1200/JCO.2003.03.059

7. D'Amico TA, Bandini LAM, Balch A, et al Quality measurement in cancer care: a review and endorsement of high-impact measures and concepts. *J Natl Compr Canc Netw*. 2020;18(3):250-259.

8. Morden NE, Chang CH, Jacobson JO, et al. End-of-life care for Medicare beneficiaries with cancer is highly intensive overall and varies widely. *Health Aff (Millwood)*. 2012;31(4):786-796. doi:10.1377/hlthaff.2011.0650

9. Goodman DC, Fisher ES, Chang CH, et al. *Quality of End-of-Life Cancer Care for Medicare Beneficiaries Regional and Hospital-Specific Analyses*. The Dartmouth Atlas;2010.

10. American Society of Clinical Oncology. QOPI-related measures. Published 2020. Accessed March 20, 2020. https://practice.asco.org/quality-improvement/quality-programs/quality-oncology-practice-initiative/qopi-related-measures

11. Krishnan M, Temel JS, Wright AA, Bernacki R, Selvaggi K, Balboni T. Predicting life expectancy in patients with advanced incurable cancer: a review. *J Support Oncol*. 2013;11(2):68-74. doi:10.12788/j.suponc.0004

12. Chow E, Harth T, Hruby G, Finkelstein J, Wu J, Danjoux C. How accurate are physicians' clinical predictions of survival and the available prognostic tools in estimating survival times in terminally ill cancer patients? a systematic review. *Clin Oncol (R Coll Radiol)*. 2001;13(3):209-218. doi:10.1007/s001740170078

13. White N, Reid F, Harris A, Harries P, Stone P. A systematic review of predictions of survival in palliative care: how accurate are clinicians and who are the experts? *PLoS One*. 2016;11(8):e0161407. doi:10.1371/journal.pone.0161407

14. Weeks JC, Cook EF, O'Day SJ, et al. Relationship between cancer patients' predictions of prognosis and their treatment preferences. *JAMA*. 1998;279(21):1709-1714. doi:10.1001/jama.279.21.1709

15. Rose JH, O'Toole EE, Dawson NV, et al. Perspectives, preferences, care practices, and outcomes among older and middle-aged patients with late-stage cancer. *J Clin Oncol*. 2004;22(24):4907-4917. doi:10.1200/JCO.2004.06.050

16. Tang ST, Chen CH, Wen FH, et al. Accurate prognostic awareness facilitates, whereas better quality of life and more anxiety symptoms hinder end-of-life care discussions: a longitudinal survey study in terminally ill cancer patients' last six months of life. *J Pain Symptom Manage*. 2018;55(4):1068-1076. doi:10.1016/j.jpainsymman.2017.12.485

17. Nipp RD, Greer JA, El-Jawahri A, et al. Coping and prognostic awareness in patients with advanced cancer. *J Clin Oncol*. 2017;35(22):2551-2557. doi:10.1200/JCO.2016.71.3404

18. Lundquist G, Rasmussen BH, Axelsson B. Information of imminent death or not: does it make a difference? *J Clin Oncol*. 2011;29(29):3927-3931. doi:10.1200/JCO.2011.34.6247

19. Zhang B, Wright AA, Huskamp HA, et al. Health care costs in the last week of life: associations with end-of-life conversations. *Arch Intern Med*. 2009;169(5):480-488. doi:10.1001/archinternmed.2008.587

20. El-Jawahri A, Traeger L, Park ER, et al. Associations among prognostic understanding, quality of life, and mood in patients with advanced cancer. *Cancer*. 2014;120(2):278-285. doi:10.1002/cncr.28369

21. Finlay E, Casarett D. Making difficult discussions easier: using prognosis to facilitate transitions to hospice. *CA Cancer J Clin*. 2009;59(4):250-263. doi:10.3322/caac.20022

22. Gensheimer MF, Henry AS, Wood DJ, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *J Natl Cancer Inst*. 2019;111(6):568-574. doi:10.1093/jnci/djy178

23. Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. *JAMA Netw Open*. 2018;1(3):e180926-e180926. doi:10.1001/jamanetworkopen.2018.0926

24. Brajer N, Cozzi B, Gao M, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA Netw Open*. 2020;3(2):e1920733-e1920733. doi:10.1001/jamanetworkopen.2019.20733

25. Parikh RB, Manz C, Chivers C, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw Open*. 2019;2(10):e1915997-e1915997. doi:10.1001/jamanetworkopen.2019.15997

26. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73. doi:10.7326/M14-0698

27. Penn Medicine. Epic clarity. Published 2020. Accessed March 20, 2020.https://www.med.upenn.edu/dac/epic-clarity-data-warehousing.html

28. Machine-generated mortality estimates and nudges to promote advance care planning discussion among cancer patients. Clinicaltrials.gov identifier: NCT03984773. Updated April 24, 2020. Accessed June 1, 2020. https://clinicaltrials.gov/ct2/show/NCT03984773

29. National Technical Information Service. Limited access death master file: final rule establishing certification program for access to death master file in effect. Published 2020. Accessed March 20, 2020. https://classic.ntis.gov/products/ssa-dmf/#

30. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Paper presented at: Proceedings of the 23rd International Conference on Machine Learning; June 25, 2006; Pittsburgh, PA.

31. Wu YC, Lee WC. Alternative performance measures for prediction models. *PLoS One*. 2014;9(3):e91249. doi:10.1371/journal.pone.0091249

32. ECOG-ACRIN Research Group. ECOG performance status. Published 2020. Accessed March 20, 2020. https://ecog-acrin.org/resources/ecog-performance-status

33. Lieffers JR, Baracos VE, Winget M, Fassbender K. A comparison of Charlson and Elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data. *Cancer*. 2011;117(9):1957-1965. doi:10.1002/cncr.25653

34. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. *Ann Intern Med*. 2014;160(2):122-131. doi:10.7326/M13-1522

35. Haslam A, Prasad V. Estimation of the percentage of US patients with cancer who are eligible for and respond to checkpoint inhibitor immunotherapy drugs. *JAMA Netw Open*. 2019;2(5):e192535-e192535. doi:10.1001/jamanetworkopen.2019.2535

36. Bertsimas D, Dunn J, Pawlowski C, et al. Applied informatics decision support tool for mortality predictions in patients with cancer. *JCO Clin Cancer Inform*. 2018;2:1-11. doi:10.1200/CCI.18.00003

37. Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *J Gen Intern Med*. 2018;33(6):921-928. doi:10.1007/s11606-018-4316-y

38. Datta SS, Ghosal N, Daruvala R, et al. How do clinicians rate patient's performance status using the ECOG performance scale? a mixed-methods exploration of variability in decision-making in oncology. *Ecancermedicalscience*. 2019;13:913. doi:10.3332/ecancer.2019.913

39. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019. doi:10.1001/jama.2019.18058

40. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

41. Emanuel EJ, Wachter RM. Artificial intelligence in health care: will the value match the hype? *JAMA*. 2019;321(23):2281-2282. doi:10.1001/jama.2019.4914

42. Manz CR, Parikh RB, Evans CN, et al. Integrating machine-generated mortality estimates and behavioral nudges to promote serious illness conversations for cancer patients: Design and methods for a stepped-wedge cluster randomized controlled trial. *Contemp Clin Trials*. 2020;90:105951. doi:10.1016/j.cct.2020.105951