

# Applied Informatics Decision Support Tool for Mortality Predictions in Patients With Cancer

original report

**Purpose** With rapidly evolving treatment options in cancer, the complexity in the clinical decision-making process for oncologists represents a growing challenge magnified by oncologists' disposition of intuition-based assessment of treatment risks and overall mortality. Given the unmet need for accurate prognostication with meaningful clinical rationale, we developed a highly interpretable prediction tool to identify patients with high mortality risk before the start of treatment regimens.

**Methods** We obtained electronic health record data between 2004 and 2014 from a large national cancer center and extracted 401 predictors, including demographics, diagnosis, gene mutations, treatment history, comorbidities, resource utilization, vital signs, and laboratory test results. We built an actionable tool using novel developments in modern machine learning to predict 60-, 90- and 180-day mortality from the start of an anticancer regimen. The model was validated in unseen data against benchmark models.

**Results** We identified 23,983 patients who initiated 46,646 anticancer treatment lines, with a median survival of 514 days. Our proposed prediction models achieved significantly higher estimation quality in unseen data (area under the curve, 0.83 to 0.86) compared with benchmark models. We identified key predictors of mortality, such as change in weight and albumin levels. The results are presented in an interactive and interpretable tool ([www.oncomortality.com](http://www.oncomortality.com)).

**Conclusion** Our fully transparent prediction model was able to distinguish with high precision between highest- and lowest-risk patients. Given the rich data available in electronic health records and advances in machine learning methods, this tool can have significant implications for value-based shared decision making at the point of care and personalized goals-of-care management to catalyze practice reforms.

Clin Cancer Inform. © 2018 by American Society of Clinical Oncology

Dimitris Bertsimas

Jack Dunn

Colin Pawlowski

John Silberholz

Alexander Weinstein

Ying Daisy Zhuo

Eddy Chen

Aymen A. Elfiky

Author affiliations and support information (if applicable) appear at the end of this article.

**Corresponding author:**

Aymen A. Elfiky, MD, MPH, MSc, MBA; Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215; e-mail: [Aymen\\_Elfiky@dfci.harvard.edu](mailto:Aymen_Elfiky@dfci.harvard.edu).

## INTRODUCTION

With the growing nuances of cancer management, the calculus of data integration and clinical intuition over time complicates physicians' decision-making capacities. Considering each new treatment line, especially within the context of noncurative-intent treatments, clinicians must account for potentially incremental benefits of a treatment against comprehensive mortality risk, given such factors as performance status, previous treatments, aggregate toxicities, and patients' evolving goals of care across their disease trajectory. Factually, physicians overestimate prognosis in cancer,<sup>1,2</sup> and patient preferences are sensitive to these estimates.<sup>3-5</sup> Organizations, including the National Quality Forum and ASCO, have identified chemotherapy

administration to patients with no clinical benefit as the most pervasive and superfluous practice in oncology.<sup>6,7</sup> Indeed, unqualified treatment of progressive disease increases symptom burden, aggregate adverse events, and costly interventions that have little morbidity or mortality benefit.<sup>8-11</sup>

Mortality prediction is a keystone of clinical decision making. [Table 1](#) lists key studies on mortality prediction in cancer.<sup>12,13</sup> These studies fall under two broad categories. Studies A1-A11 predicted mortality using detailed patient information in relatively small patient cohorts. Studies B1-B3 used the large-scale SEER registry database with clinical characteristics at the time of diagnoses and treatments. Reported areas under the curve (AUC) range from 0.71 to 0.87. In recent years, some prediction models have

been made accessible as online tools, which has allowed the results from prognostic research to be readily available for use at the point of care. Although such tools offer potential convenience to users, the underlying data and models limit the potential for higher prediction quality while maintaining interpretability. Most tools are based on cancer registry cohorts and are limited to a small number of characteristics at the time of diagnosis, overlooking critical information from clinical trajectories with implications for patients' prognoses. Even with more detailed data sets, a tool interface requires learning accurate models with few variable inputs. Black-box models, such as artificial neural networks (B1) or gradient boosted trees that rely on large numbers of predictors, would not be suitable for such interactive tools. Furthermore, those methods offer little explanation to physicians regarding why such predictions were made. These practical points highlight critical concerns that explain the limited use of available prognostic tools in daily clinical workflows.

In our study, we used patient electronic health record (EHR) data from a tertiary cancer center and a novel team-constructed algorithm to develop a predictive tool that estimates the probability of mortality for an individual patient being proposed their next treatment. The tool's practical value lies in being:

1. Personalized: The tool uses EHR inputs for patients, their cancer, and the proposed treatment to output a mortality risk adjusted for these characteristics.
2. Clinically cogent: Physicians easily understand the reasoning behind the algorithm's outputs through interpretable decision trees that illustrate key variable predictors.
3. Evidence and data driven: The tool is informed by over 23,000 patient records with  $\geq 400$  predictors drawn from demographics, comorbidities, treatments, laboratory values, and genomic results, and so on.
4. Actionable: Clinicians can compare between proposed treatments for a patient with respect to mortality risk estimates to make informed recommendations and decisions.
5. Validated and accurate: The accuracy and AUC in unseen patient data show

encouraging results compared with competing approaches.

6. On the basis of novel machine learning: The methodology leverages two optimization algorithms developed by our team: (1) predictive classification trees,<sup>29</sup> and (b) predictive statistical methods for missing data imputation.<sup>30</sup>

The following sections detail our approach to developing the predictive model and tool.

## METHODS

### Patients

We retrospectively obtained patient data from the EHR and linked Social Security Administration mortality data for patients at the Dana-Farber/Brigham and Women's Cancer Center from 2004 through 2014. Study eligibility required that patients be at least 18 years old at cancer diagnosis and have received at least one anticancer treatment over the course of their care. The primary outcomes were mortality rates at 60, 90 and 180 days after initiation of an anticancer regimen, including chemotherapy, immunotherapy, and targeted therapy. If the patient's date of death was missing and the last known date alive was before the cutoff, that patient's record was censored for predicting that outcome. Each observation corresponded to a patient initiating an anticancer regimen, which was systematically recorded in the EHR.

### Data

We considered 401 predictive features, including demographics, diagnoses, treatments, vital signs, laboratory results, and so on (Data Supplement). Missing values were imputed using the optimal-impute algorithm,<sup>30</sup> which frames the imputation task as a family of optimization problems and solves directly. We chose this algorithm because the data had many key predictors that had a large number of observations missing, a scenario in which traditional methods, such as complete-case analysis and mean imputation, often give poor predictions in the end, whereas this algorithm had demonstrated significant improvement in downstream prediction tasks compared with classic methods. The quality of imputation was evaluated in a sensitivity analysis against alternative methods. We used treatment lines initiated from 2004 to 2011 as the training

**Table 1.** Summary of Mortality Prediction Studies in Patients With Advanced Cancer

Study	Model	Population	Training Sample Size, No.	Covariable Size, No.	Data	Model AUC	First Author
A1	Palliative Prognostic Score, exponential multiple regression	Terminal cancer	519	36, selected 6	Demographics, cancer status and treatment, symptoms, laboratory values	Not reported	Pirovano <sup>14</sup>
A2	Palliative Performance Index, multiple regression	Terminal cancer	150	25, selected 5	Demographics, cancer status, symptoms	Not reported	Morita <sup>15</sup>
A3	Memorial Sloan Kettering Cancer Center nomograms, accelerated failure time model	Metastatic prostate cancer after castration	409	7 total	Demographics, symptoms, laboratory values	Not reported	Smaletz <sup>16</sup>
A4	Cancer Prognostic Score, Cox regression model	Terminal cancer	356	26, selected 8	Demographics, cancer status and treatment, symptoms	Not reported	Chuang <sup>17</sup>
A5	Intrahospital Cancer Mortality Risk Model, multivariable logistic regression	Hospitalized patients with cancer	334	14, selected 5	Demographics, cancer status and treatment, symptoms, laboratory values	0.82 (intrahospital mortality)	Bozcuk <sup>18</sup>
A6	Glasgow Prognostic Score, Cox regression model	Inoperable non-small-cell lung cancer	161	10 total	Demographics, cancer status and treatment, symptoms, laboratory values	Not reported	Forrest <sup>19</sup>
A7	Objective Prognostic Score, Cox regression model	Terminal cancer	209	17, selected 7	Demographics, cancer status and treatment, symptoms, laboratory values	Not reported	Suh <sup>20</sup>
A8	Artificial neural networks	Non-small-cell lung cancer	440	Thousands of variables, selected 5	Demographics, cancer status and treatment, gene expression	Not reported	Chen <sup>21</sup>
A9	Support vector machine ensemble	All cancers	869	18 total	Demographics, cancer status (including receptor status for patients with breast cancer)	0.76 (2-year survival), 0.80 (1-year survival), 0.87 (6-month survival)	Gupta <sup>22</sup>
A10	Bayesian network	Breast cancer	78	7 clinical, 232 genes	Demographics, cancer status, gene expression	0.85 (predicting good prognosis)	Gevaert <sup>23</sup>
A11	Imminent Mortality Predictor in Advanced Cancer study, Cox regression model	Hospital-admitted patients with cancer with advanced solid tumors	468	13 PCA-derived weights and 22 static variables; selected 1 PCA-derived weight and 5 static variables	Rothman Index <sup>24</sup> time-series trajectory, and clinicopathologic variables	0.736 (30-day mort.), 0.722 (60-day mort.), 0.710 (90-day mort.), 0.717 (180-day mort.)	Adelson <sup>25</sup>

(Continued on following page)

**Table 1.** Summary of Mortality Prediction Studies in Patients With Advanced Cancer (Continued)

Study	Model	Population	Training Sample Size, No.	Covariable Size, No.	Data	Model AUC	First Author
B1	Artificial neural networks	Breast and colorectal cancers	5,169 (breast), 5,007 (colorectal)	21 (breast), 32 (colorectal)	Demographics, cancer status and treatment, symptoms	0.78/0.87 (5-year survival: breast/colorectal)	Burke <sup>26</sup>
B2	Graph-based semisupervised learning and others	Breast cancer	40,000	16 total	Demographics, cancer status and treatment	0.81 (5-year survival)	Kim <sup>27</sup>
B3	Decision trees and others	Breast cancer	202,932	16 total	Demographics, cancer status and treatment	Not reported	Delen <sup>28</sup>

Abbreviations: AUC, area under the curve; mort., mortality; PCA, principal component analysis.

set and assessed each model's predictive performance using those initiated from 2012 to 2014 as the validation set. The institutional review boards of Dana-Farber Cancer Institute/Partners Healthcare and the Massachusetts Institute of Technology approved this study.

### Model and Tool Development

The mortality predictions are based on a novel decision tree algorithm, optimal classification trees.<sup>29</sup> We selected decision trees for their advantage in interpretability, where the predictions are based on a few decision splits on variables of high importance. Such tree structures can readily model nonlinearities and interactions between variables. However, classic decision tree methods typically cannot achieve the same level of accuracy as their less interpretable counterparts, such as artificial neural networks and gradient boosted trees. To mitigate the trade-off between interpretability and prediction accuracy, we made use of the optimal classification trees algorithm that trains a single decision tree with state-of-the-art performance, achieving high accuracy without the need to sacrifice interpretability.

During the training process, we tuned the parameters to maximize performance on a separate holdout set to avoid overfitting. After training the model, for interpretation and verification, we generated the following: (1) a decision tree visualization, whereby experts could trace the algorithm's logic and predict mortality risks to corroborate clinical relevance; and (2) feature importance scores, which provided an estimate of the relative importance of key variables in the mortality predictions.

We next built the interactive tool for physicians on the basis of the trained decision trees. The tool is made available as a Web-based application ([www.oncomortality.com](http://www.oncomortality.com)) in the format of a patient characteristics questionnaire. The clinician user has the option to select 60-, 90-, or 180-day mortality predictions and cancer types (all cancers, breast, lung, ovarian, and so on.).

The user is then prompted to answer a short series of adaptive questions corresponding to the decision splits in the trained decision tree, until the tool ultimately generates a predicted mortality risk specific to the patient at hand. Examples are provided in Results.

### Performance Comparisons

We evaluated the performance of optimal classification trees in unseen patient data for the quality of mortality predictions. To demonstrate our improvement in prognostic quality, we sought to ideally compare our model against established prognostic studies on the same patient populations. For such comparisons, the outcome variables (short-term mortality) needed to be aligned, and the predictor variables from prior studies needed to be available. Unfortunately, among the identified studies, no such comparison could be conducted. As proxies, a variety of other machine learning models were trained and validated on the same set of data for comparisons (comparative models detailed in the Data Supplement). We report the prediction accuracy (number of correctly classified positive and negative samples over total samples) on the basis of a default 50% threshold and the positive predictive value (PPV), a clinically relevant measure, at a fixed sensitivity level of 0.6. For further

assessment of the model at varying thresholds, we report the AUC for each model in addition to plotting the receiver operating characteristic curves and the PPV against sensitivity for various thresholds. The analyses were repeated for each of the three time horizons (60-, 90-, and 180-day mortality) and subgrouped by cancer type.

## RESULTS

### Patient and Treatment Characteristics

A total of 23,983 patients were selected in the cohort, spanning a spectrum of breast (26.7%), lung (14.9%), ovarian (7.7%), colorectal (5.9%), and other solid tumors and hematologic malignancies. Among those, 14,427 (60.2%) were in the training set and 9,556 (39.8%) were in the validation set. These patients initiated 46,646 total treatment lines; 2,619 (5.6%) of new treatment line starts were followed by the patient's death within 60 days, and 44,027 (94.4%) were not. Baseline characteristics of groups are presented in the Data Supplement. Overall, new treatment lines that led to mortality were associated with more malignant cancers (eg, lung and pancreatic cancer) and heavier prior resource utilization. Patients in this group on average also had higher disease burden and staging, as well as a higher number of prescribed medications, inpatient/outpatient visits, and blood infusions. The treatment lines that led to mortality were also associated with more comorbid conditions, such as congestive heart failure, stroke, diabetes. Finally, the laboratory test results for those lines that led to mortality were often significantly

worse than the ones that did not lead to mortality (lower weight and albumin, higher tumor markers, suppressed blood counts, and so on.). The median survival was 514 days for all patients.

### Interpretable Tool on the Basis of Machine Learning

We trained the optimal classification trees to predict the 60-, 90-, and 180-day mortality. The model produced predictions with accuracies of 94.9%, 93.3%, and 86.1% at a 50% threshold, and AUCs of 0.86, 0.84, and 0.83, respectively. We further trained the model to predict the mortality for each subgroup of cancer sites, achieving similarly high estimation qualities, with AUCs ranging from 0.77 to 0.90 (Data Supplement).

On the basis of the prediction algorithm, we developed the tool, which is available online (screenshot depicted in Fig 1). Once the clinician enters the desired time horizon and cancer type, his or her answers to adaptive questions concludes with the tool output of a risk estimation. In this example, in response to the 60-day mortality risk for a patient with lung cancer, the tool adaptively prompts the clinician to input the percentage change in weight and the albumin levels. With these two input values, the tool presents a final risk for 60-day mortality of 46.79%.

### Model Interpretation

Each model trained with optimal classification trees presents a highly interpretable decision tree on which the tool is based. A tree stratifies patients into risk groups on the basis of values of a sequence of key variables, the selection of which is learned automatically by the model. As an example, Figure 2 presents the tree that predicts the 60-day mortality for patients with breast cancer. The total number of metastatic solid tumors is the first splitting variable, patients with few metastatic solid tumors (left of the tree), are placed into the lower-risk branch (0.4%) compared with the higher-risk one (5.84%). Within each branch, patients are further stratified on the basis of other variables. For example, in the lower-risk branch, if the pulse is above 86 beats per minute, the patient has a mortality risk of 1.31%; if the pulse is below 86 beats per minute, depending on the age (younger or older than 60 years), the risk is 0% and 0.88%, respectively. In the higher-risk branch, many of the decisions

**Fig 1.** Screenshot of the cancer mortality prediction questionnaire. The clinician enters responses to a few questions regarding the patient's medical history, many of which can be automatically populated via electronic health record integration, and the tool will immediately generate predictions. Available online at [www.oncomortality.com](http://www.oncomortality.com).

I would like to predict my patient's risk of:

60-day mortality

Site:

LUNG

What is the patient's change in weight as a percentage from its average over the past 90 days? 4416 PATIENTS (9.78%)

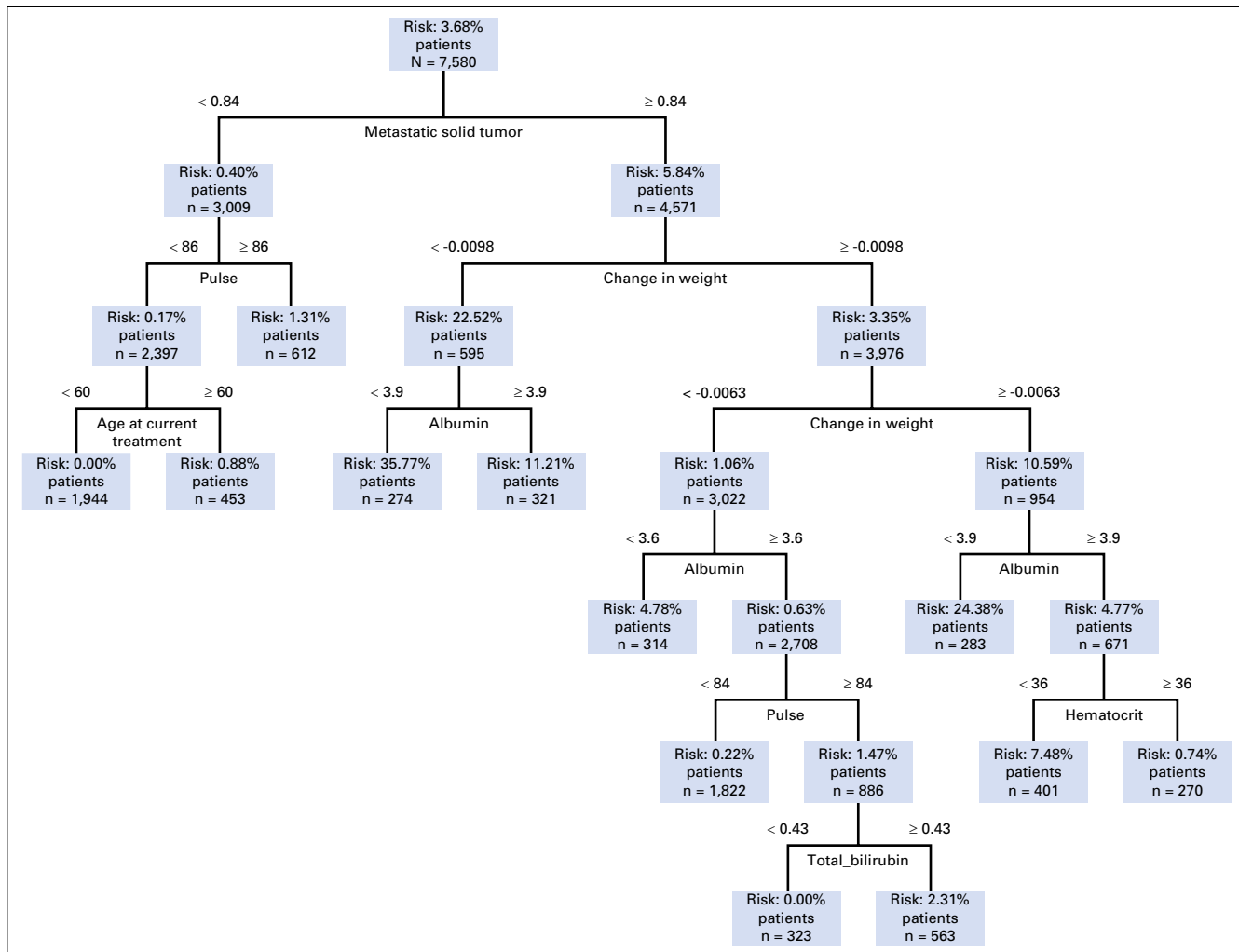
-0.01

What is the patient's albumin level in g/dL? 1059 PATIENTS (25.59%)

1.9

Final risk estimation: 156 PATIENTS

46.79%



**Fig 2.** Mortality prediction tree for patients with breast cancer for 60-day mortality. Patients are stratified on the basis of a sequence of variables and are eventually placed into a mortality risk bin.

are based on change in weight, albumin, and other laboratory test results.

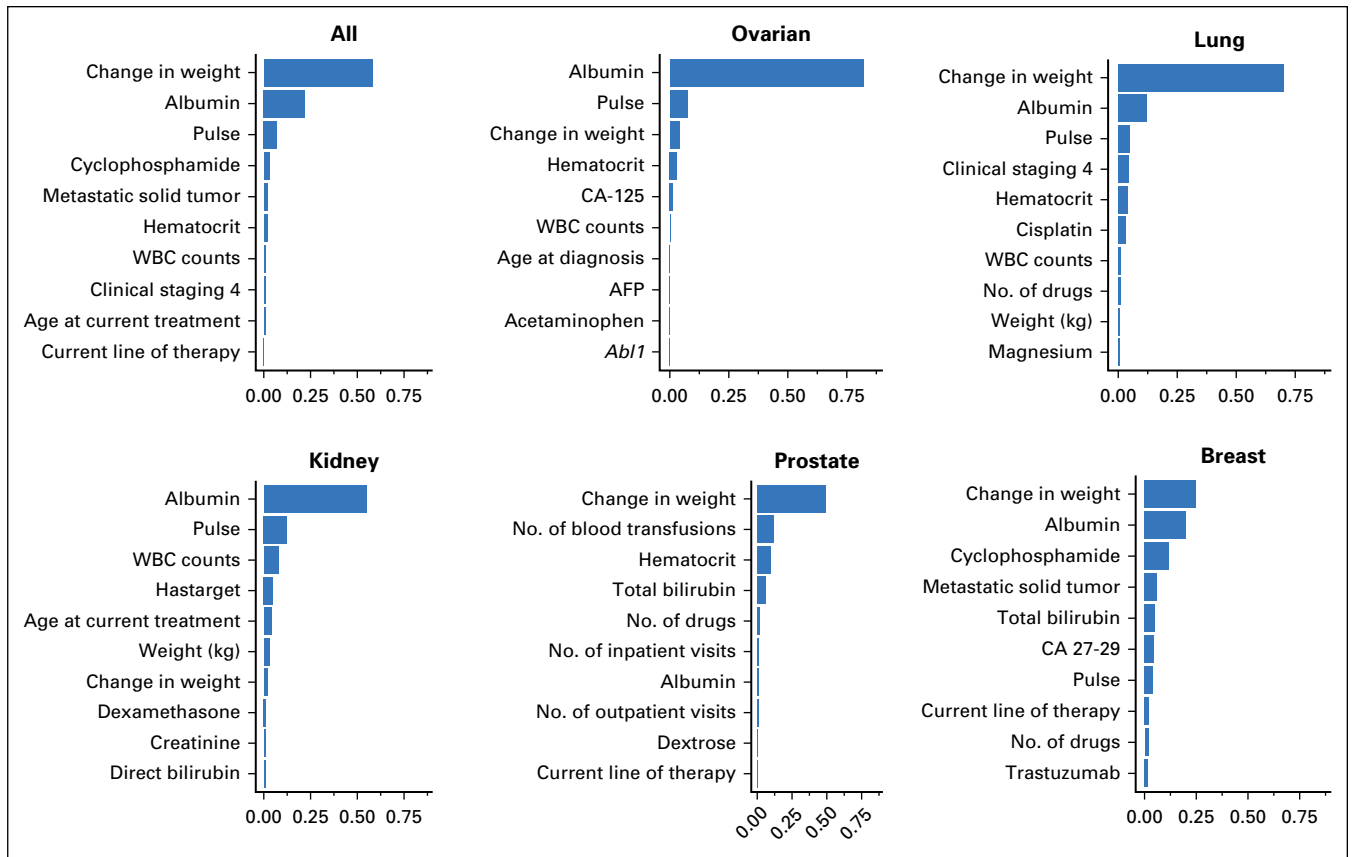
This model also produces feature importance rankings in mortality prediction (Fig 3). In tree-based models, the feature importance score measures the relative contribution of a particular feature in the model on the basis of the frequency this variable is selected for splitting and the improvement in model performance at each split; the score of all variables sum up to 1.<sup>29</sup> Among all patients, percentage change in weight from a patient's moving average over the past 90 days is the most important predictive feature of mortality (a more drastic decrease is associated with higher risk of mortality). Albumin level, pulse, WBC counts, total bilirubin, and weight were the next predictive variables.

Although the feature importance does not demonstrate the interaction across predictive variables, the relationship was evidently characterized by the

tree structure. As an example, the Data Supplement depicts a case where the change in weight interacts with albumin in the predictions of mortality.

### Machine Learning Models Comparison of Performance

Among all the transparent machine learning models being compared against, optimal classification trees achieved the best performance in validation (Table 2). The only method that improves over optimal classification trees is the black-box method of gradient boosted trees. The receiver operating characteristic curves and PPV against sensitivity curves for each of the methods (Fig 4) both showed consistently strong predictive power across all thresholds at all time horizons. In the subgroup analyses, we found a similar performance for our method (detailed model comparison results



**Fig 3.** Feature importance in 60-day mortality prediction for patients (all cancer and by cancer sites). The importance score is based on the relative contribution to the model performance of each feature during the optimal classification trees training process. The 10 most important predictors are shown in this figure. *Ab1*, Abelson murine leukemia viral oncogene homolog 1 gene (ABL proto-oncogene 1); AFP, alpha-fetoprotein; CA, cancer antigen.

and subgroup breakdowns provided in the Data Supplement).

### DISCUSSION

Within the context of the growing momentum toward value-based health care delivery,<sup>6</sup> we developed a clinically actionable prognostic tool for individual mortality prediction among patients with cancer before initiating a systemic treatment. Developed with the intent to augment physicians' clinical decision making at the point of care, our tool necessarily accounts across the entirety of a patient's cancer treatment journey to provide personalized prognostic insight. Moreover, such insight can address an unmet need by catalyzing the proper engagement of a physician, a patient, and his or her caregiver(s) in a personal, objectively assessed reflection on the implications of a proposed treatment in relation to the overall goals of care.

Our predictive tool uses novel machine learning methods to support the movement toward fully personalized, evidence-based treatment decisions. The prediction model's accuracy draws

from the combination of (1) a large cohort of general patients with cancer, (2) longitudinal EHR data that provide extensive and nuanced information compared with registry and claims data, (3) novel machine learning methods with high interpretability and state-of-the-art performance, and (4) curated clinical intuition and substantiated knowledge from experienced oncologists on the team that facilitated the model class selection and feature engineering process. The optimal classification trees model we used is fully transparent and interpretable, and produces highly accurate results. More importantly, decision trees closely resemble a human approach to decision making and, as a result, our machine learning-based approach to medical decision support is consistent with physicians' mental process of identifying patterns from experience, but doing so using a much broader and more representative cohort of patient base. Because of the transparency and interpretability inherent in our model, as technology continues to develop and more data are collected, the model will improve while maintaining its resemblance to physicians' decision-making processes.

**Table 2.** Model Performances (accuracy at 50% threshold, PPV at sensitivity level of 0.6, and AUC) for 60-Day, 90-Day, and 180-Day Mortality Predictions in the Validation Set, Comparing Optimal Classification Trees Against Five Other Prediction Models

Model Performance	60-Day Mortality	90-Day Mortality	180-Day Mortality
Accuracy (%)			
Logistic regression (fewer predictors)	94.6	93.0	83.4
Logistic regression	94.3	92.8	84.4
Regularized logistic regression	94.9	93.1	84.4
CART decision tree	94.0	92.0	81.8
Optimal classification tree	94.9	93.3	86.1
Gradient boosted trees	95.0	93.6	87.2
Positive predictive value at sensitivity level of 0.6 (%)			
Logistic regression (fewer predictors)	11.3	15.9	32.2
Logistic regression	14.0	18.2	35.9
Regularized logistic regression	16.1	21.4	37.5
CART decision tree	20.2	21.0	39.2
Optimal classification tree	20.2	27.5	43.1
Gradient boosted trees	28.2	34.7	53.4
AUC			
Logistic regression (fewer predictors)	0.74	0.76	0.76
Logistic regression	0.73	0.74	0.75
Regularized logistic regression	0.79	0.80	0.79
CART decision tree	0.81	0.81	0.79
Optimal classification tree	0.86	0.84	0.83
Gradient boosted trees	0.90	0.89	0.87

Abbreviations: AUC, area under the curve; CART, classification and regression trees; PPV, positive predictive value.

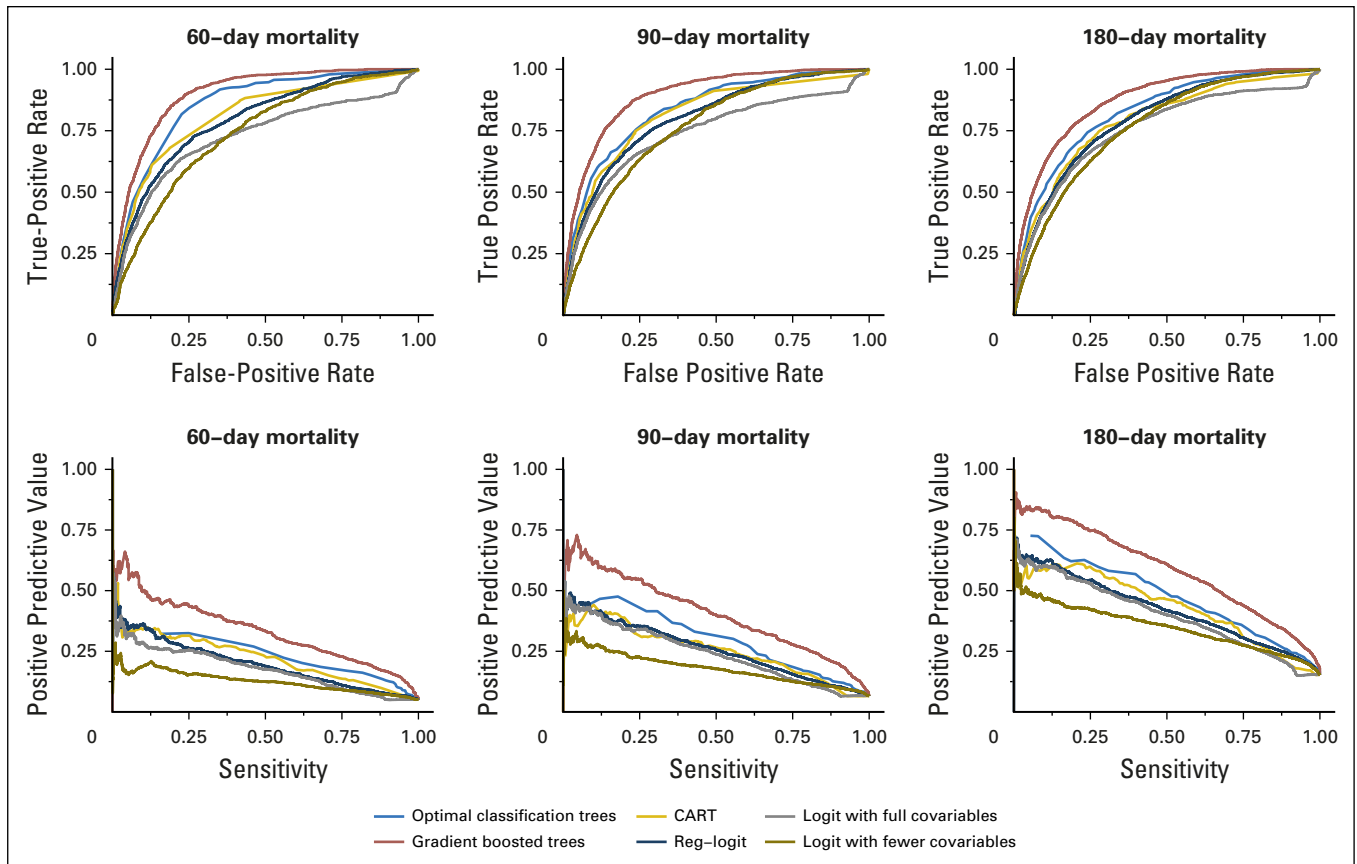
Operationally, given the complexity of knowledge management in evolving precision oncology care practices,<sup>31</sup> we envision our tool's adaptive information curation system seamlessly integrating with EHRs as a value-driven component of a chemotherapy-order-entry workflow to augment clinical decision making and informed consent. Because curated information is automatically populated from data repositories, physicians verify populated inputs and acknowledge mortality risk outputs transpiring within real time. On the back end, the novel optimal-impute<sup>30</sup> method fulfills the necessary preprocessing characteristic of a nimble data platform that facilitates accurate results in the sensitive downstream tasks of mortality prediction and knowledge-based practices.<sup>31</sup>

The rich data constitute the other key reason for our high prediction quality. Because this study predicts mortality in a large population with available EHR data, we were able to study a much broader set of covariables than typical mortality prediction studies. We studied 401 covariables in total, including 289 variables

encoding gene mutation results, 52 on recent treatments, and 18 on recent laboratory and vital test results. We acknowledge that because of the sparsity of data collection on gene mutation results, these variables were not selected by the models, but as more genomic data are collected consistently in the future, we believe these predictors will play a more important role and further improve the models. The longitudinal nature of the EHR data used in this study further enables the study of patient characteristics through time. For instance, we included the percentage change in weight from the 90-day moving average weight measurement, a predictor not used in previous mortality prediction work.

Performing head-to-head comparisons against existing prognostic studies and online tools is of interest. However, among identified short-term mortality studies, differences in patient populations and subjective clinical variables that our data do not contain make definitive comparison of results difficult. Nonetheless, the IMPAC (Imminent Mortality Predictor for Advanced





**Fig 4.** Receiver operating characteristic curves and positive predictive value versus sensitivity plots for 60-day, 90-day, and 180-day mortality predictions, comparing the following methods: optimal classification trees, CART (classification and regression trees), logistic regression with fewer variables, logistic regression with all variables, regularized logistic regression, and gradient boosted trees.

Cancer<sup>25</sup> study’s comparative metrics of mortality prediction specifically within the hospital setting are notable for their improved accuracy, thereby highlighting potential predictive benefits to analyzing carefully curated patient cohorts and care episodes. Significantly, both studies achieve high levels of predictive performance that provide a benchmark for future studies across the cancer care continuum.

Our study has several limitations. As a single-institution retrospective study, it is subject to data selection and measurement biases. In addition, validation was performed using an internal data set, which underscores our plans to test prospectively in an external cohort. Looking forward, despite the continuing importance of chemotherapies in the cancer treatment arsenal, we already anticipate the need to refactor our model in light of novel drugs and changing treatment paradigms. Nonetheless, disparities in care resources and access to treatments persist in certain geodemographic and sociodemographic areas where our tool may hold relevance. Encouragingly, as precision-based molecular and genomic sequencing data

are collected in more patients along the disease trajectory, our model and performance are expected to improve. To this end, the iterative process of the model’s learning, calibration, and validation is essential to its prognostic performance and continued relevance to clinical practice. Further prospective studies with this tool are being planned to demonstrate the clinical utility of this tool as part of a platform measuring the impact on informed consent engagement, mortality, and patient experience metrics, including hospital length of stay, earlier hospice enrollment, and quality of life.

In conclusion, we developed a clinically actionable tool for mortality prediction for patients with cancer initiating systemic therapy. Such capability leveraging machine learning, temporally nuanced data, and clinical expertise presents the opportunity for value capture at point of care.

DOI: <https://doi.org/10.1200/CCI.18.00003>

Published online on [ascopubs.org/journal/cci](https://ascopubs.org/journal/cci) on June 7, 2018.

#### AUTHOR CONTRIBUTIONS

**Conception and design:** Dimitris Bertsimas, Alexander Weinstein, Aymen A. Elfiky  
**Financial support:** Aymen A. Elfiky  
**Administrative support:** Ying Daisy Zhuo  
**Collection and assembly of data:** Ying Daisy Zhuo, Aymen A. Elfiky  
**Data analysis and interpretation:** All authors  
**Manuscript writing:** All authors  
**Final approval of manuscript:** All authors  
**Accountable for all aspects of the work:** All authors

#### AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to [www.asco.org/rwc](http://www.asco.org/rwc) or [ascopubs.org/jco/site/ifc](http://ascopubs.org/jco/site/ifc).

#### **Dimitris Bertsimas**

No relationship to disclose

#### **Jack Dunn**

No relationship to disclose

#### **Colin Pawlowski**

No relationship to disclose

#### **John Silberholz**

No relationship to disclose

#### **Alexander Weinstein**

**Employment:** Celgene/Juno (I)

**Stock and Other Ownership Interests:** Celgene (I)

#### **Ying Daisy Zhuo**

No relationship to disclose

#### **Eddy Chen**

No relationship to disclose

#### **Aymen A. Elfiky**

No relationship to disclose

#### **Affiliations**

**Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, John Silberholz, Alexander Weinstein, and Ying Daisy Zhuo,** Massachusetts Institute of Technology, Cambridge; **Eddy Chen,** Massachusetts General Hospital Cancer Center; Harvard Medical School; **Aymen A. Elfiky,** Dana-Farber Cancer Institute; Brigham and Women's Hospital; Harvard Medical School, Boston, MA.

#### **Support**

Supported by funding from the Dana-Farber Cancer Institute (A.A.E.) and a National Science Foundation Predoctoral Fellowship (C.P.).

#### **REFERENCES**

1. Glare P, Virik K, Jones M, et al: A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ* 327:195-198, 2003
2. Stone PC, Lund S: Predicting prognosis in patients with advanced cancer. *Ann Oncol* 18:971-976, 2007
3. Brundage MD, Davidson JR, Mackillop WJ: Trading treatment toxicity for survival in locally advanced non-small cell lung cancer. *J Clin Oncol* 15:330-340, 1997
4. Silvestri G, Pritchard R, Welch HG: Preferences for chemotherapy in patients with advanced non-small cell lung cancer: Descriptive study based on scripted interviews. *BMJ* 317:771-775, 1998
5. Hirose T, Yamaoka T, Ohnishi T, et al: Patient willingness to undergo chemotherapy and thoracic radiotherapy for locally advanced non-small cell lung cancer. *Psychooncology* 18:483-489, 2009
6. Schnipper LE, Smith TJ, Raghavan D, et al: American Society of Clinical Oncology identifies five key opportunities to improve care and reduce costs: The top five list for oncology. *J Clin Oncol* 30:1715-1724, 2012
7. National Quality Forum: Endorsement summary: Cancer measures. [https://www.qualityforum.org/News\\_And\\_Resources/Endorsement\\_Summaries/Cancer\\_Measures\\_Endorsement\\_Summary.aspx](https://www.qualityforum.org/News_And_Resources/Endorsement_Summaries/Cancer_Measures_Endorsement_Summary.aspx)
8. Wright AA, Zhang B, Keating NL, et al: Associations between palliative chemotherapy and adult cancer patients' end of life care and place of death: Prospective cohort study. *BMJ* 348:g1219, 2014
9. de Gramont A, Figer A, Seymour M, et al: Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. *J Clin Oncol* 18:2938-2947, 2000

10. Emanuel EJ, Young-Xu Y, Levinsky NG, et al: Chemotherapy use among Medicare beneficiaries at the end of life. *Ann Intern Med* 138:639-643, 2003
11. Greer JA, Pirl WF, Jackson VA, et al: Effect of early palliative care on chemotherapy use and end-of-life care in patients with metastatic non-small-cell lung cancer. *J Clin Oncol* 30:394-400, 2012
12. Glare P, Sinclair C, Downing M, et al: Predicting survival in patients with advanced disease. *Eur J Cancer* 44:1146-1156, 2008
13. Kourou K, Exarchos TP, Exarchos KP, et al: Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 13:8-17, 2014
14. Pirovano M, Maltoni M, Nanni O, et al: A new palliative prognostic score: A first step for the staging of terminally ill cancer patients. *J Pain Symptom Manage* 17:231-239, 1999
15. Morita T, Tsunoda J, Inoue S, et al: Survival prediction of terminally ill cancer patients by clinical symptoms: Development of a simple indicator. *Jpn J Clin Oncol* 29:156-159, 1999
16. Smaletz O, Scher HI, Small EJ, et al: Nomogram for overall survival of patients with progressive metastatic prostate cancer after castration. *J Clin Oncol* 20:3972-3982, 2002
17. Chuang RB, Hu WY, Chiu TY, et al: Prediction of survival in terminal cancer patients in Taiwan: Constructing a prognostic scale. *J Pain Symptom Manage* 28:115-122, 2004
18. Bozcuk H, Koyuncu E, Yildiz M, et al: A simple and accurate prediction model to estimate the intrahospital mortality risk of hospitalised cancer patients. *Int J Clin Pract* 58:1014-1019, 2004
19. Forrest LM, McMillan DC, McArdle CS, et al: Evaluation of cumulative prognostic scores based on the systemic inflammatory response in patients with inoperable non-small-cell lung cancer. *Br J Cancer* 89:1028-1030, 2003
20. Suh SY, Choi YS, Shim JY, et al: Construction of a new, objective prognostic score for terminally ill cancer patients: A multicenter study. *Support Care Cancer* 18:151-157, 2010
21. Chen YC, Ke WC, Chiu HW: Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med* 48:1-7, 2014
22. Gupta S, Tran T, Luo W, et al: Machine-learning prediction of cancer survival: A retrospective study using electronic administrative records and a cancer registry. *BMJ Open* 4:e004007, 2014
23. Gevaert O, De Smet F, Timmerman D, et al: Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22:e184-e190, 2006
24. Rothman MJ, Rothman SI, Beals J IV: Development and validation of a continuous measure of patient condition using the electronic medical record. *J Biomed Inform* 46:837-848, 2013
25. Adelson K, Lee DKK, Velji S, et al: Development of Imminent Mortality Predictor for Advanced Cancer (IMPAC), a tool to predict short-term mortality in hospitalized patients with advanced cancer. *J Oncol Pract* 14:e168-e175, 2018
26. Burke HB, Goodman PH, Rosen DB, et al: Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 79:857-862, 1997
27. Kim J, Shin H: Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *J Am Med Inform Assoc* 20:613-618, 2013
28. Delen D, Walker G, Kadam A: Predicting breast cancer survivability: A comparison of three data mining methods. *Artif Intell Med* 34:113-127, 2005
29. Bertsimas D, Dunn J: Optimal classification trees. *Mach Learn* 106:1039-1082, 2017
30. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: An optimization approach. *J Mach Learn Res* (in press)
31. Elfiky A, Zhang D, Krishnan Nair HK: Practice innovation: The need for nimble data platforms to implement precision oncology care. *Discov Med* 20:27-32, 2015