Journal of the American Medical Informatics Association, 28(3), 2021, 605–615
doi: 10.1093/jamia/ocaa254
Advance Access Publication Date: 1 December 2020
Research and Applications

Research and Applications

Development of a "meta-model" to address missing data, predict patient-specific cancer survival and provide a foundation for clinical decision support

Jason M. Baron[1,2,*,†], Ketan Paranjape[3], Tara Love[4], Vishakha Sharma[4], Denise Heaney[3,*], and Matthew Prime[5,*]

[1]Independent Consultant, (Somerville, MA) on Behalf of Roche Diagnostics Corporation, Indianapolis, Indiana, USA, [2]Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA,**[3]Roche Diagnostics Corporation, North America, Indianapolis, Indiana, [4]Roche Diagnostics Corporation, Santa Clara, California, USA and [5]Roche Diagnostics Corporation, Riehen, Basel Stadt, Switzerland

*These authors contributed equally.

[†]Jason Baron performed his portion of the work on this manuscript in his capacity as an independent consultant on behalf of Roche Diagnostics and not within his academic roles the Massachusetts General Hospital and Harvard Medical School.
Corresponding Author: Jason M. Baron, MD, 7 Maxwell's Green, Somerville, MA 02144, USA; jason.baron@contractors.-roche.com

Received 19 June 2020; Revised 27 August 2020; Editorial Decision 15 September 2020; Accepted 27 October 2020

ABSTRACT

Objective: Like most real-world data, electronic health record (EHR)–derived data from oncology patients typically exhibits wide interpatient variability in terms of available data elements. This interpatient variability leads to missing data and can present critical challenges in developing and implementing predictive models to underlie clinical decision support for patient-specific oncology care. Here, we sought to develop a novel ensemble approach to addressing missing data that we term the "meta-model" and apply the meta-model to patient-specific cancer prognosis.
Materials and Methods: Using real-world data, we developed a suite of individual random survival forest models to predict survival in patients with advanced lung cancer, colorectal cancer, and breast cancer. Individual models varied by the predictor data used. We combined models for each cancer type into a meta-model that predicted survival for each patient using a weighted mean of the individual models for which the patient had all requisite predictors.
Results: The meta-model significantly outperformed many of the individual models and performed similarly to the best performing individual models. Comparisons of the meta-model to a more traditional imputation-based method of addressing missing data supported the meta-model's utility.
Conclusions: We developed a novel machine learning–based strategy to underlie clinical decision support and predict survival in cancer patients, despite missing data. The meta-model may more generally provide a tool for addressing missing data across a variety of clinical prediction problems. Moreover, the meta-model may address other challenges in clinical predictive modeling including model extensibility and integration of predictive algorithms trained across different institutions and datasets.

Key words: missing data, imputation, clinical decision support, meta-model, machine learning, survival

INTRODUCTION

Background and significance

Predictive models trained using real-world clinical data offer tremendous potential to provide patients and their clinicians patient-specific information regarding diagnosis, prognosis, or optimal therapeutic course.[1–10] For example, a recent high-profile study trained a machine learning model using hundreds of thousands of patient records to forecast the development of acute kidney injury.[9] However, key challenges have limited the introduction of machine learning–based predictive models into real clinical settings.[3] One set of challenges relates to interpatient variability in data availability. In most real-world datasets, many patients will lack recorded findings for many clinical factors.[3,7,11–13] For example, some hospitals may have a laboratory test menu that includes more than 1000 unique orderable tests. Most patients will have had at most a small fraction of these possible tests. A similar pattern involving substantial "missing data" would usually be observed for nonlaboratory clinical data, including other diagnostic studies, elements of patient history, and physical exam findings. The issue of data heterogeneity becomes particularly significant when considering time series data; even patients who have similar diagnostic tests or physical exam maneuvers performed may have them at different time points or repeated at varying intervals.[3]

Many commonly used machine learning algorithms require complete datasets and cannot directly use for training or prediction datasets containing missing data. Data scientists commonly employ several strategies to enable use of real-world data that include missing elements in predictive analyses. One strategy involves preprocessing a set of clinical data by "imputing" missing data elements.[14] While there are numerous variations on imputation and related approaches, including single imputation, multiple imputation, and expectation maximization, most imputation approaches are fundamentally designed to use available data to estimate the distribution or value of each element of missing data.[7,11–13,15–18] The preprocessed dataset, including both actual and imputed clinical findings, can then be used to train standard machine learning models or can be applied to trained models to generate predictions. However, imputation, while very useful in many contexts has important limitations. Most imputation algorithms assume that data are missing at random (MAR); because diagnostic studies are selected and ordered in response to the clinical setting, most clinical datasets will violate the MAR assumption.[3,11,18] Likewise, imputation can introduce additional uncertainty and inaccuracy into predictions and may obscure some of the intuition behind some predictive models.

As described subsequently, we propose and demonstrate an alternative approach to imputation in addressing missing data. We term this new approach the meta-model. To develop and apply the meta-model, we consider the problem of patient-specific prognosis prediction in patients with advanced oncologic disease. While population-based survival statistics are available across a wide range of cancer types and patients, patient-specific information can be harder to discern. For example, based on national SEER (Surveillance, Epidemiology, and End Results) statistics, the overall 5-year survival of patients with stage IV colon cancer is just 14%.[19] However, some individual patients will have a considerably better than average survival. The critical question for an oncologist then, when seeing an individual patient, is not the population survival, but rather what the individual patient's prognosis is. Individualizing patient prognosis is not itself a new endeavor. On the contrary, numerous published studies describe clinical risk factors that portend better or worse prognosis. For example, prior studies clearly establish that patients with colon cancer experience shorter survival on average if they have comorbid diabetes.[20] While a clinician may take these types of published findings into account when considering prognosis, their true clinical utility can be quite limited. In particular, patients may have multiple clinical factors that individually could convey improved or worsened prognosis; there would usually not be a viable strategy to calculate the aggregate impact of these multiple factors. Indeed, prior studies have shown limitations of the human brain in manually making predictions based on a large number of predictors.[2] Thus, as a secondary focus of this article, we propose, validate, and demonstrate a strategy to apply machine learning to the development of patient-specific Kaplan-Meier survival curves. These patient-specific curves may offer oncologists and other clinicians the opportunity to more accurately assess patient prognosis and communicate risk to patients.

Objectives

This article has 2 objectives. The primary objective is to develop and demonstrate a novel meta-model approach to addressing missing data. As described in detail subsequently, our meta-model concept includes an ensemble of underlying models based on varying predictors with the final output based on an aggregate of all individual models for which a patient has complete data. The meta-model may also address other challenges in predictive clinical decision support (CDS) implementation including model extensibility and integration of predictive algorithms trained across different institutions and datasets.

The secondary objective of this article is to develop a method for generating patient-specific Kaplan-Meier survival curves.

MATERIALS AND METHODS

An overview of our methods is shown as Figure 1. Using clinical data from patients with metastatic colorectal cancer (CRC), metastatic breast cancer, and advanced lung cancer, we first developed a set of survival prediction models intended to individualize patient prognosis. After developing and validating the individual models, we combined the individual models for each dataset into a meta-model. We demonstrate that this meta-model can provide an alternative to imputation in addressing missing data and may offer several key advantages. Key methodologic points are described subsequently with additional detail provided in the Supplementary Methods.

Patient cohorts

We defined patient cohorts from 3 subsets of the nationwide Flatiron Health electronic health record–derived de-identified database[21]: (1) metastatic CRC, (2) advanced non-small cell lung cancer, and (3) metastatic breast cancer. For each patient in our cohort, we extracted and assembled outcome data (time surviving after the date of advanced diagnosis) and selected clinical features that were commonly available and that we thought might help to predict prognosis ("predictors"). We randomly split the cases for each cohort into a training and a testing partition in an approximately 80:20 ratio.

Survival outcome

We captured survival, defined as the number of days between advanced tumor diagnosis and death for use as our outcome variable. Patients were censored to the time of their last encounter.

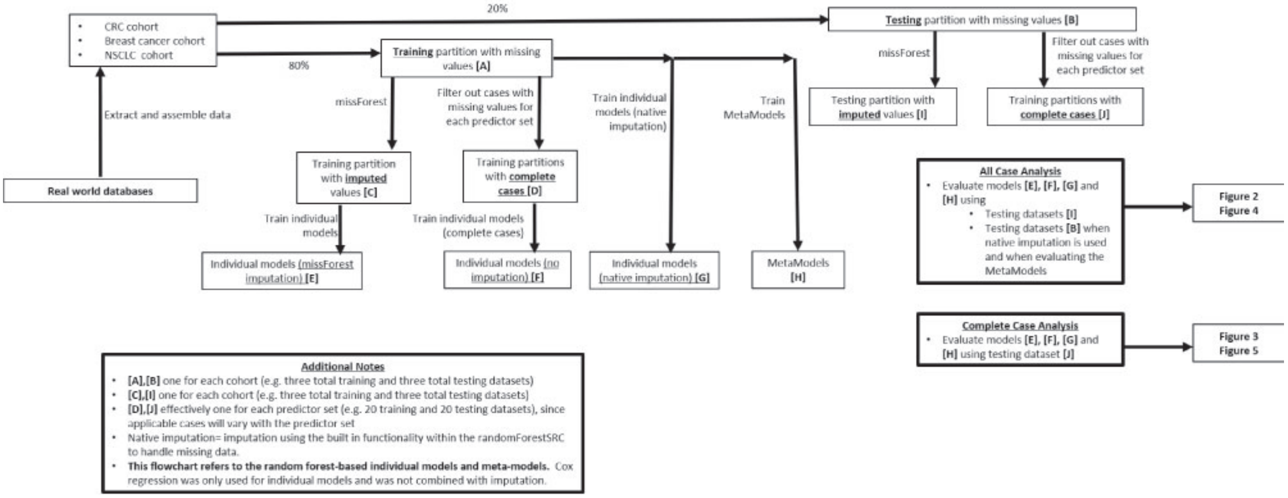

CRC cohort
Breast cancer cohort
NSCLC cohort
20%
80%
Extract and assemble data
Real world databases
Training partition with missing values [A]
missForest
Filter out cases with missing values for each predictor set
Train individual models (native imputation)
Train MetaModels
Training partition with imputed values [C]
Training partitions with complete cases [D]
Train individual models
Train individual models (complete cases)
Individual models (missForest imputation) [E]
Individual models (no imputation) [F]
Individual models (native imputation) [G]
MetaModels [H]
Testing partition with missing values [B]
missForest
Filter out cases with missing values for each predictor set
Testing partition with imputed values [I]
Training partitions with complete cases [J]
All Case Analysis
Evaluate models [E], [F], [G] and [H] using
Testing datasets [I]
Testing datasets [B] when native imputation is used and when evaluating the MetaModels
Figure 2
Figure 4
Complete Case Analysis
Evaluate models [E], [F], [G] and [H] using testing dataset [J]
Figure 3
Figure 5
Additional Notes
[A],[B] one for each cohort (e.g. three total training and three total testing datasets)
[C],[I] one for each cohort (e.g. three total training and three total testing datasets)
[D],[J] effectively one for each predictor set (e.g. 20 training and 20 testing datasets), since applicable cases will vary with the predictor set
Native imputation= imputation using the built in functionality within the randomForestSRC to handle missing data.
This flowchart refers to the random forest-based individual models and meta-models. Cox regression was only used for individual models and was not combined with imputation.

Figure 1. Overview of data, model development, and performance assessment. Shown is an overview of the approach used to extract, analyze, and model the data and assess model performance. Details are described in the Materials and Methods section and in the Supplementary Methods. CRC: colorectal cancer; NSCLC: non-small cell lung cancer.

Predictors and predictor sets

We considered patient demographics, tumor characteristics, molecular biomarkers, and laboratory test results for use as predictors (Table 1). We prioritized potential predictors for inclusion based on factors including data availability and expected predictive value. That is, we favored predictors available on a larger number of patients (based on a preliminary exploration of the data) and those that, based on our domain expertise, we thought were more likely to be of value. In addition, we considered the usability of predictors both in our analysis and in potential downstream applications (eg, we preferred predictors often available in structured form with meanings that are substantially standardized across institutions).

We further grouped the predictors into "predictor sets" (Table 1). Similar to how we selected the predictors themselves, we selected the predictor set groupings based on patterns of data availability (eg, we preferentially included lab tests commonly performed together in the same predictor set) with an emphasis on developing predictor sets for which many or all patients would have all of the available predictors. The number of total cases along with an accounting of censored patients and deceased patients, available for use with each predictor set, is shown as Table 2.

Individual model development

For each cohort and set of predictors, we trained 2 different survival models: one linear Cox regression model and another based on a random forest. We used the R survival package[22] to develop and validate the linear models and the R randomForestSRC package[23–26] to develop the tree-based models. We included 100 trees per individual model. Additional detail on these models is available in the Supplementary Methods.

Addressing missing data in training and testing individual models

We used 3 strategies to train and test individual models in the setting of missing data. The first strategy involved using complete cases with respect to each predictor set (ie, patients were excluded from training or testing who did not have all of the requisite predictors needed). The second strategy involved imputation. We imputed missing predictor values using the random forest–based imputation algorithm, missForest as described in greater detail in the Supplementary Methods. The third strategy to addressing missing data involved leveraging built-in functionality to handle missing data within the randomforestSRC package[23–26]; we term this third strategy native imputation.

Meta-model conceptual approach

Aiming to integrate survival predictions across individual models, improve overall prediction accuracy, and offer other important practical properties, as described subsequently, we developed an approach we termed the meta-model. Conceptually, the meta-model (Supplementary Figure S1) starts by training individual prediction models, such as the individual survival models described previously. It then assigns a weight to each model, based on the model's accuracy, such that models that tend to predict outcomes with greater accuracy are more heavily weighted (specific approach to assigned weights described subsequently). The meta-model can then be applied to test patients by computing all individual models for which the patient has the necessary predictors and then taking a weighted average of the predictions produced by these individual models. We note that the meta-model in part represents an adaptation of Breiman's[27] stacked regression.

Development of the survival meta-model

We trained 1 meta-model for each cohort (3 meta-models total) using the predictor sets shown in Table 1. For each predictor set, we trained a random forest–based model using methods paralleling the development of the individual survival prediction models. (We refer to each of these individual random forest models within the meta-model as an "individual model.") To assign a weight to each individual model, we performed 5-fold cross validation of each individual model capturing the median cross-validation area under the receiver-operating characteristic curve (AUROC) of the model at 500, 1000, and 1500 days. We then transformed these median AUROC values into model weights using 1 of several weighting

Table 1. Predictors used in each model

		Advanced Lung Cancer								Metastatic Breast Cancer					Metastatic Colorectal Cancer						
Variable	Type	A	B	C	D	E	F	G	H	A	B	C	D	E	A	B	C	D	E	F	G
Gender	Categorical	1	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
GroupStage	Categorical	0	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
dxageYrs	Numerical	1	0	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
CrcSite	Categorical	x	x	x	x	x	x	x	x	x	x	x	x	x	0	1	1	1	1	1	1
Histology	Categorical	0	1	1	1	0	1	1	1	x	x	x	x	x	x	x	x	x	x	x	x
SmokingStatus	Categorical	0	1	1	1	0	1	1	1	x	x	x	x	x	x	x	x	x	x	x	x
leukocytes	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
hemoglobin	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
platelets	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
hematocrit	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
erythrocytes	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
creatinine	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
lymphocytes	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	1	1
protein	Lab	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	1	0	0
bilirubin	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	1	1
urea.nitrogen	Lab	x	x	x	x	x	x	x	x	x	x	x	x	x	0	0	0	1	1	1	1
carcinoembryonic.ag	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	0	0	0	0	1	1	1
calcium	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	0	0
sodium	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	0	0
potassium	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	0	0
alkaline.phosphatase	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	0	0
lymphocytes.per.100.leukocytes	Lab	x	x	x	x	x	x	x	x	0	0	0	0	1	0	0	0	1	1	0	0
monocytes.per.100.leukocytes	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x	x	x	x	x
carbon.dioxide	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x	x	x	x	x
monocytes	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x	x	x	x	x
chloride	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x	x	x	x	x
lactate.dehydrogenase	Lab	0	0	0	0	0	0	0	1	x	x	x	x	x	x	x	x	x	x	x	x
ER	Categorical	x	x	x	x	x	x	x	x	0	0	0	1	1	0	0	0	0	0	0	0
PR	Categorical	x	x	x	x	x	x	x	x	0	0	0	1	1	0	0	0	0	0	0	0
HER2	Categorical	x	x	x	x	x	x	x	x	0	0	0	1	1	0	0	0	0	0	0	0
EGFR	Categorical	0	0	0	0	1	1	1	1	x	x	x	x	x	x	x	x	x	x	x	x
ALK	Categorical	0	0	0	0	1	1	1	1	x	x	x	x	x	x	x	x	x	x	x	x
KRAS	Categorical	x	x	x	x	x	x	x	x	x	x	x	x	x	0	0	0	0	0	1	1
glucose	Lab	x	x	x	x	x	x	x	x	x	x	x	x	x	0	0	0	1	1	0	0
BRAF	Categorical	x	x	x	x	x	x	x	x	x	x	x	x	x	0	0	0	0	0	0	1
Random number	Numerical	0	0	0	0	0	0	0	0	x	x	x	x	x	x	x	x	x	x	x	x

1→ predictor is included in specified predictor set; 0→ predictor is not included in specified predictor set, but is included in other predictor sets for the cancer type; x→ predictor is not included in any predictor sets within the corresponding cancer type.

Table 2. Patient characteristics by predictor set

		Complete Cases			
Cohort	Predictor Set	n	%	Total Deceased	Total Censored
Advanced lung cancer	A	6558	100	1146	5412
	B	6559	100	1146	5413
	C	6558	100	1146	5412
	D	4479	68	639	3840
	E	3254	50	577	2677
	F	3253	50	577	2676
	G	2285	35	338	1947
	H	157	2	21	136
Metastatic breast cancer	A	5045	100	1633	3412
	B	5045	100	1633	3412
	C	5045	100	1633	3412
	D	4795	95	1548	3247
	E	2854	57	818	2036
Metastatic colorectal cancer	A	6742	100	1760	4982
	B	6743	100	1761	4982
	C	6742	100	1760	4982
	D	3888	58	888	3000
	E	3435	51	768	2667
	F	2759	41	588	2171
	G	1163	17	269	894

functions having the form $w = (x - 0.5)^n$, where w is the weight assigned to the model prediction, x represents the median cross-validation AUROC, and n represents an exponent (see **Supplementary Methods** for additional detail). A value of n = 2 was used for all analyses unless otherwise specified.

Model evaluation metrics

We evaluated our models by comparing predicted survival to actual survival for patients in the test partition. We primarily used AUROC, describing each models ability to discriminate patients alive vs deceased at various time points. We specifically considered AUROC at 500, 1000, and 1500 days post advanced diagnosis. We calculated AUROC using the method described by Heagerty et al.[28] as implemented in the R package survivalROC.[28,29] In addition, as described in greater detail in the **Supplementary Methods**, we evaluated model calibration by comparing the actual deaths with the predicted deaths for groups of patients over various time windows.

RESULTS

As described in the methods and Figure 1, we tested our models in 2 settings: (1) complete case analysis (we used only test cases with all needed predictors available) and (2) all case analysis (we used test cases regardless of predictor data availability).

Complete case analysis

As shown in Figure 2, we first considered the performance of individual models in comparison to the meta-model when applied to complete cases within the test data. Although the meta-model is intended for application to patients with a wide range of predictor data availability, for comparison purposes, we applied the meta-model in this analysis to the same subsets of test data used for the complete case evaluation of each individual models. Figure 2 also includes corresponding individual Cox regression models for comparison.

The best-performing individual models achieved an AUROC > 0.7 in predicting mortality at 500, 1000, and 1500 days. In almost all cases, the meta-models outperformed the individual models on comparable datasets and in some cases achieved AUROCs approaching or exceeding 0.8. The difference in performance between individual models and the meta-models was most pronounced when considering the simplest individual models (eg, the “A” and “B” models); this makes sense, given that the meta-models in many patients would have been able to leverage a much wider array of predictors than the simple individual models. More interestingly, the meta-models also seem to modestly outperform the more complex individual models in most cases. **Supplementary Table S1** provides additional training and testing AUROC values.

All case analysis

Because an important goal of the meta-models was applicability to patients with a wide range of predictor data availability, we also compared the meta-model with the individual models when applied to all test patients. For this analysis, we imputed missing test data for application to the individual models. Because the meta-models were designed to accommodate variability in predictor data availability, no imputation was used with the meta-models; however, all test patients were used to evaluate both the individual models and the meta-models.

Area under the receiver-operating characteristic curve

Model performance in the all case analysis is shown in Figure 3 (for AUROCs at 1000 days) and in **Supplementary Figures S2 and S3** (for AUROCs at 500 and 1500 days, respectively). As shown, the meta-model significantly outperforms many of the individual models and performs similarly to the best performing individual models. **Supplementary Table S1** provides additional training and testing AUROC values.

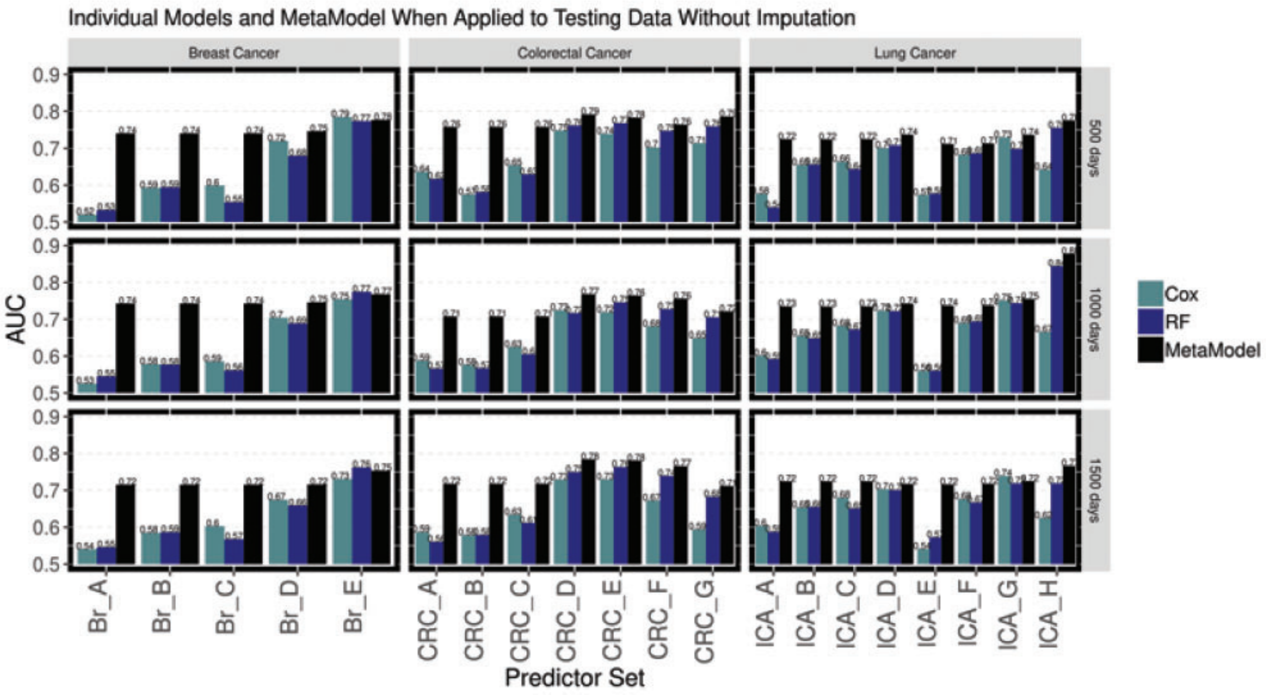


Figure 2. Model area under the receiver-operating characteristic curve (AUC when each model is applied to a subset of patients in the test partition having data available for all of the predictors in the corresponding predictor set. For example, consider the set of bars above "CRC_D" on the x-axis. The individual Cox regression model and random forest (RF) series then denote AUC values for the individual Cox and RF models corresponding to colorectal cancer (CRC) predictor set D; the meta-model bar represents the CRC meta-model. All 3 bars above CRC_D are based on a subset of the patients in the test partition who have data available for all the predictors included in CRC predictor set D (same subset of test patients used for all 3 bars). As shown, across all patient subsets, the meta-model performs better than or similar to the individual models. Br_: breast cancer predictor set; lCA_: lung cancer predictor set.

Impact of weighting functions

We compared various meta-model weighting functions (Supplementary Figure S4) and found that the specific weighting-function (ie, value of the exponent n) makes little difference in performance.

Model calibration

To assess model calibration, we compared predicted mortality to actual mortality for groups of patients over various time windows (Figure 4). To evaluate further the calibration of each survival model, we fit Poisson regression models with predicted mortality (log transformed) for patient-time groups as the independent variable and actual mortality as the dependent variable. Meta-models produced slopes close to 1 (range, 0.98-0.99) (Figure 4), suggesting that predicted and actual mortality agree well on average (a slope of 1 would suggest perfect alignment on average). Likewise, the individual models when used with imputation generally showed good calibration, but several cases exhibited slopes considerably further from 1 (range for individual models, 0.94-1.06). Supplementary Table S2 expands on the analysis in Figure 4 by explicitly considering whether the models are systematically over- or underestimating patient risk in lower- and higher-risk patients.

Individual patient survival

To illustrate how the meta-model might be used in practice, we plotted individual Kaplan-Meier survival curves for 9 selected patients within the test partition of the CRC dataset (Figure 5).

DISCUSSION

In this study, we demonstrate the utility both of a meta-model approach to addressing missing data and of the use of machine learning–based models to predict patient survival in advanced CRC, lung cancer, and breast cancer. We show that a meta-model method integrating a suite of underlying models using varied predictors may provide a practical strategy to accommodate missing data. With further validation, we anticipate that the meta-model method described here could be adapted to a wide range of prediction problems, spanning well beyond oncology survival prediction.

The meta-model approach is well suited to the development of clinical decision support, which was our primary aim in undertaking this work. Indeed, as shown in Figure 6, we aim to build an "app" to provide clinicians with access to patient-specific Kaplan-Meier curves. In addition to addressing missing data, the meta-model may provide a framework for interinstitutional predictive model development. For example, the meta-model could combine individual models trained using separate data sources, even at different institutions. While using multi-institutional data (as opposed to single-site data) to train predictive models would often be scientifically desirable in ensuring generalizability and in obtaining data from a sufficient number of patients, administrative challenges to data sharing outside individual health systems often make multi-institutional datasets impractical or impossible to obtain.[3] However, the meta-model approach may help to address this challenge to the extent that building a multi-institutional meta-model would only require institutions to share trained underlying individual models and not actual patient data. Moreover, we envision future vendor-developed

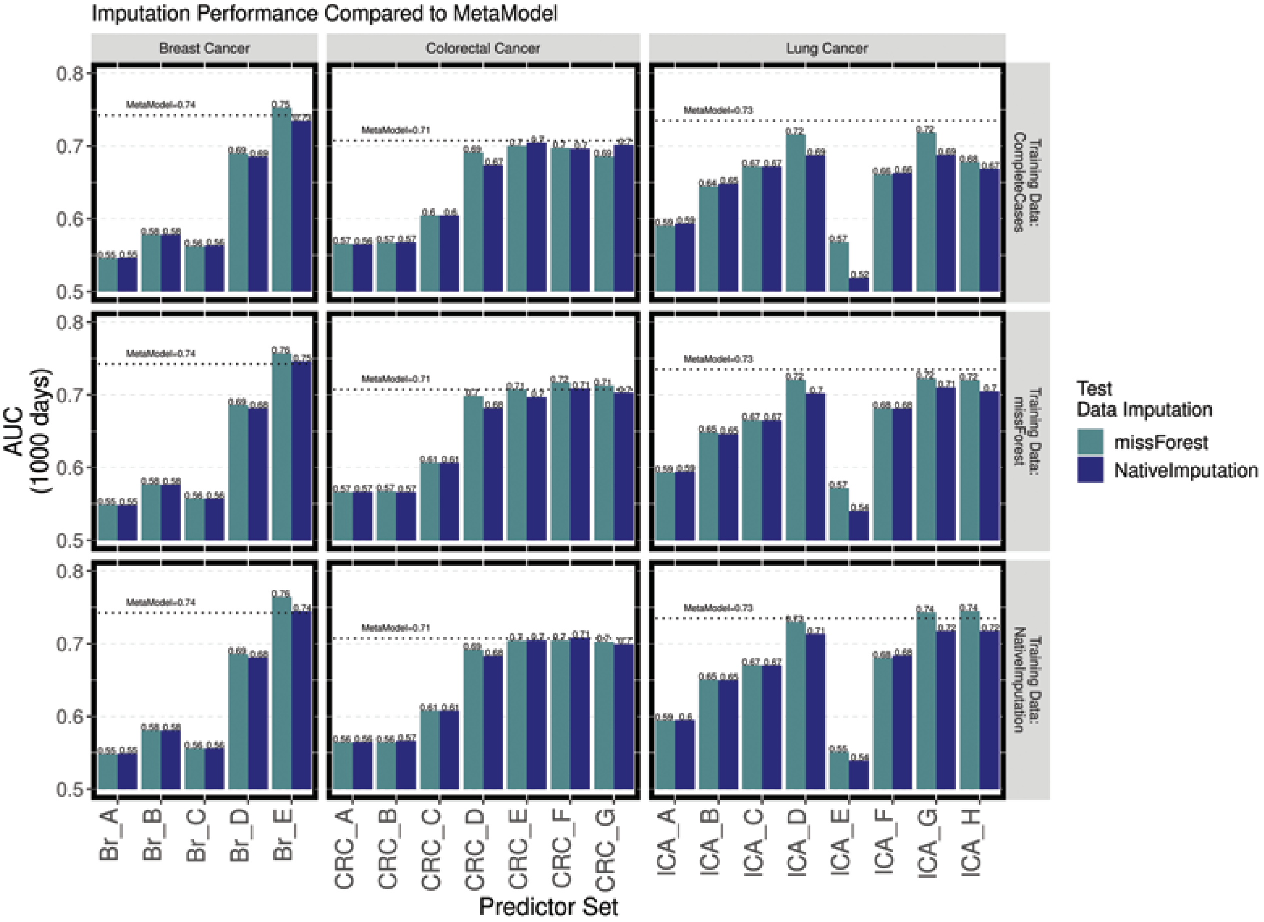


Figure 3. Model area under the receiver-operating characteristic curve (AUC) when applied to all cases from the test partition. Shown is the AUC (1000 days) of each individual model, with missing test data addressed using either missForest imputation or imputation within the random survival forest algorithm ("NativeImputation"). For comparison, also shown is the AUC of the corresponding meta-model (dashed line) when applied to the complete set of test patients. All patients within the test partition, regardless of predictor data availability, are included in all analyses. For this analysis, we trained individual models using either complete cases within the training data (top row) or addressed missing training data using missForest (middle row) or native imputation (bottom row). As shown, the meta-model in most cases outperforms the individual models; in some cases, meta-model performance is similar to or negligibly worse than individual models. Analogous figures showing AUC at 500 and at 1500 days are provided as Supplementary Figures S2 and S3. Br_: breast cancer predictor set; CRC_: CRC predictor set; ICA_: lung cancer predictor set.

CDS systems that include both a set of "starter" models as well as functionality for sites to train additional models; the systems could then combine the starter with the locally trained models using the meta-model approach described here. Finally, the meta-model approach may also be useful in identifying tests that were not performed but which could substantially reduce prognostic uncertainty (ie, tests needed for the better performing underlying models). CDS could recommend the clinician order such tests.

As noted in the introduction, most imputation and other methods for addressing missing data, including the missForest method considered here, assume data are MAR. Traditional imputation approaches may be subject to bias when data are not missing at random. For example, consider a hypothetical analysis in which patients with high values for the tumor marker carcinoembryonic antigen (CEA) are more likely to have CEA testing performed. In this case, the "observed" distribution of CEA values (ie, measured CEA results) will be higher than the unobserved distribution (ie, what the CEA results would have been in patients who did not have CEA testing). Further, suppose in this hypothetical example that high CEA correlates with poor prognosis. In this case, imputation might be prone to impute CEA results that are biased high (more in line with the distribution of observed values), and thus a survival prediction algorithm relying on these biased high imputed CEA results might be prone to overly pessimistic prognostic projections in patients without CEA testing. While a formal theoretical evaluation of the extent to which non-random missing data may bias the meta-model is beyond the scope of this article, we postulate that in many cases, the meta-model should be comparatively robust to violations of the MAR assumption. For example, consider what might happen if a meta-model approach were used in the hypothetical CEA scenario noted previously. In this case, we might expect individual models that do include CEA to on average predict poorer prognosis; however, this would be subject to adjustment based on the actual CEA result in these models. Likewise, the individual models that do not include CEA may on average provide overly pessimistic predictions in patients who do not have CEA (this might be similar to the case of imputation) and overly optimistic predictions in those who do. However, these presumably should average out at the popula-

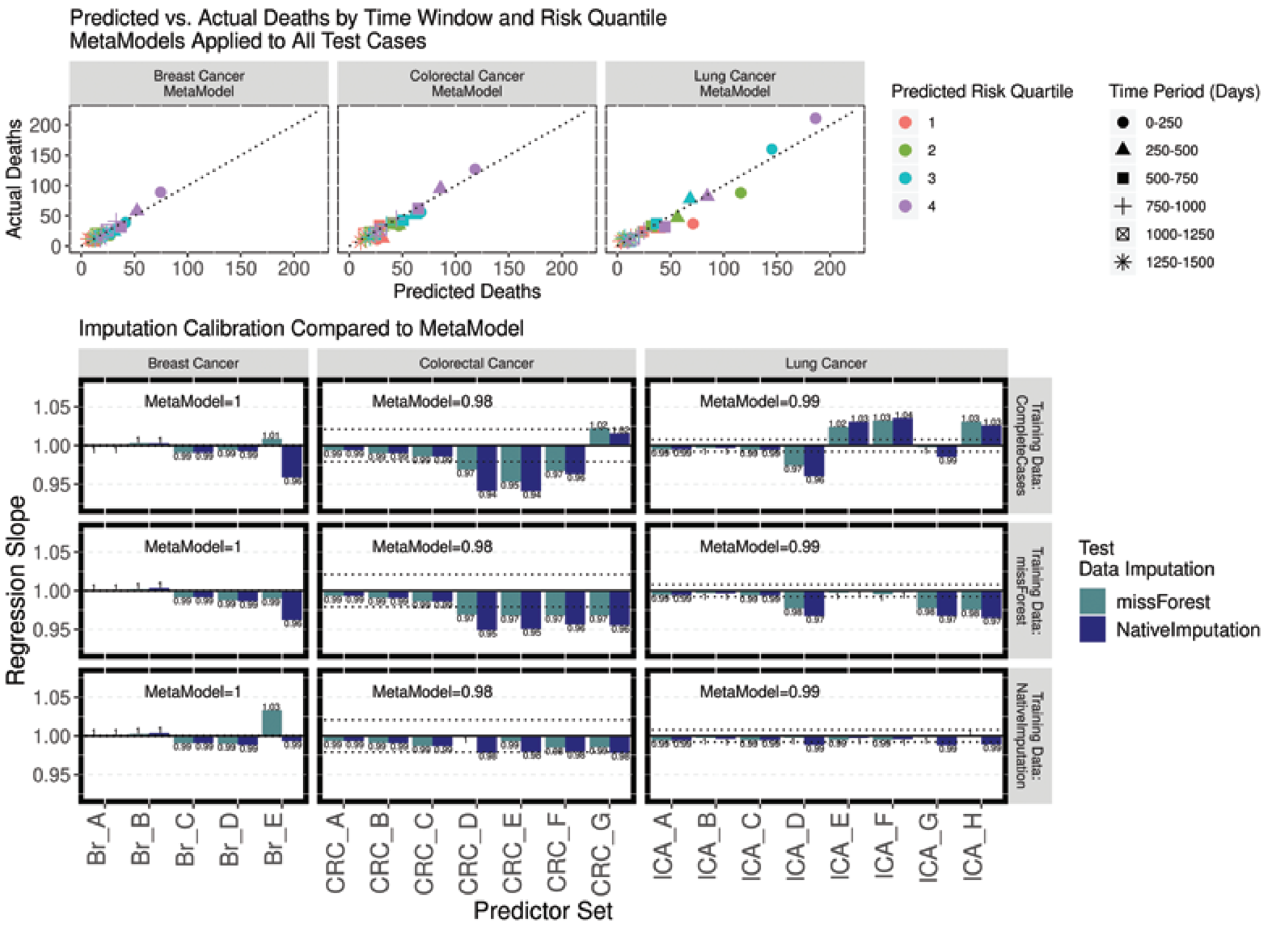


Figure 4. Model calibration. To test model calibration, we calculated each test patient's predicted mortality over 250-day time windows. We further grouped patients into risk quartiles (4 = highest risk of dying; 1 = lowest risk) and calculated the aggregate predicted mortality for each patient group–time window combination. We then compared predicted mortality to actual mortality. The scatterplots (top) plot actual vs predicted mortality for each meta-model. The dashed 45-degree line represents perfect calibration. To summarize each model's calibration, we calculated the slope of a Poisson regression line fitting actual mortality as a function of (log-transformed) predicted mortality with an intercept through the origin. Slopes of 1 would indicate perfect calibration, while slopes substantially different from 1 would indicate miscalibration. The bar graphs (bottom) plot the calibration slopes for individual models. The horizontal dashed lines in the graphs represent the meta-model calibration slope (and its mirror image around 1). Although the time windows and risk quartiles are not explicitly displayed in this bottom summary plot, the data are nonetheless grouped by risk quartile and time window, paralleling the upper scatterplots. Br_: breast cancer predictor set; CRC_: CRC predictor set; lCA_: lung cancer predictor set.

tion level if the patterns of missingness in the training and testing partition are the same. Moreover, because models with CEA would be weighted more (if they perform better), patients with CEA may tend to have predictions that overall substantially adjust for CEA results. (This, in theory, could introduce a net calibration error across the dataset). Future research will be needed to evaluate whether our speculation regarding the robustness of the meta-model to the MAR assumptions holds in practice; however, the calibration data provided here (Figure 4) empirically may be supportive. A useful area for future work would be to investigate explicitly whether patterns of data missingness impact model calibration.

While we had hypothesized that the meta-model would provide better performance than traditional methods of addressing missing data, in our experimental comparison, the meta-model did not universally outperform imputation. When applied to the complete set of test patients and assessed in terms of AUROC, imputation in combination with the best performing individual models performed similar to and in some cases, slightly better than the meta-model applied to the same patients. The meta-model may in some cases be more robust to calibration errors introduced due to data missing not at random (see Figure 4 and the paragraph preceding this one). Nonetheless, given the comparable performance of the approach, coupled with improved transparency of the meta-model and the practical applications noted previously, we expect that the meta-model will provide a useful tool. Future work will be needed to generalize our assessment of the meta-model to a range of prediction problems. Likewise, we selected only 2 imputation methods for comparison; we selected missForest in part because it had been shown to work well for laboratory test results in prior research[7,12] and because it can impute both numerical and categorical variables, but additional work comparing the meta-model to additional imputation methods could be informative.

We were surprised that the specific weighting function used to aggregate the predictions from the underlying models had little impact on overall meta-model performance. We had expected that higher values of the exponent n, which weight better-performing underlying models more heavily, would have led to better overall performance. While we do not have a full explanation, we expect that

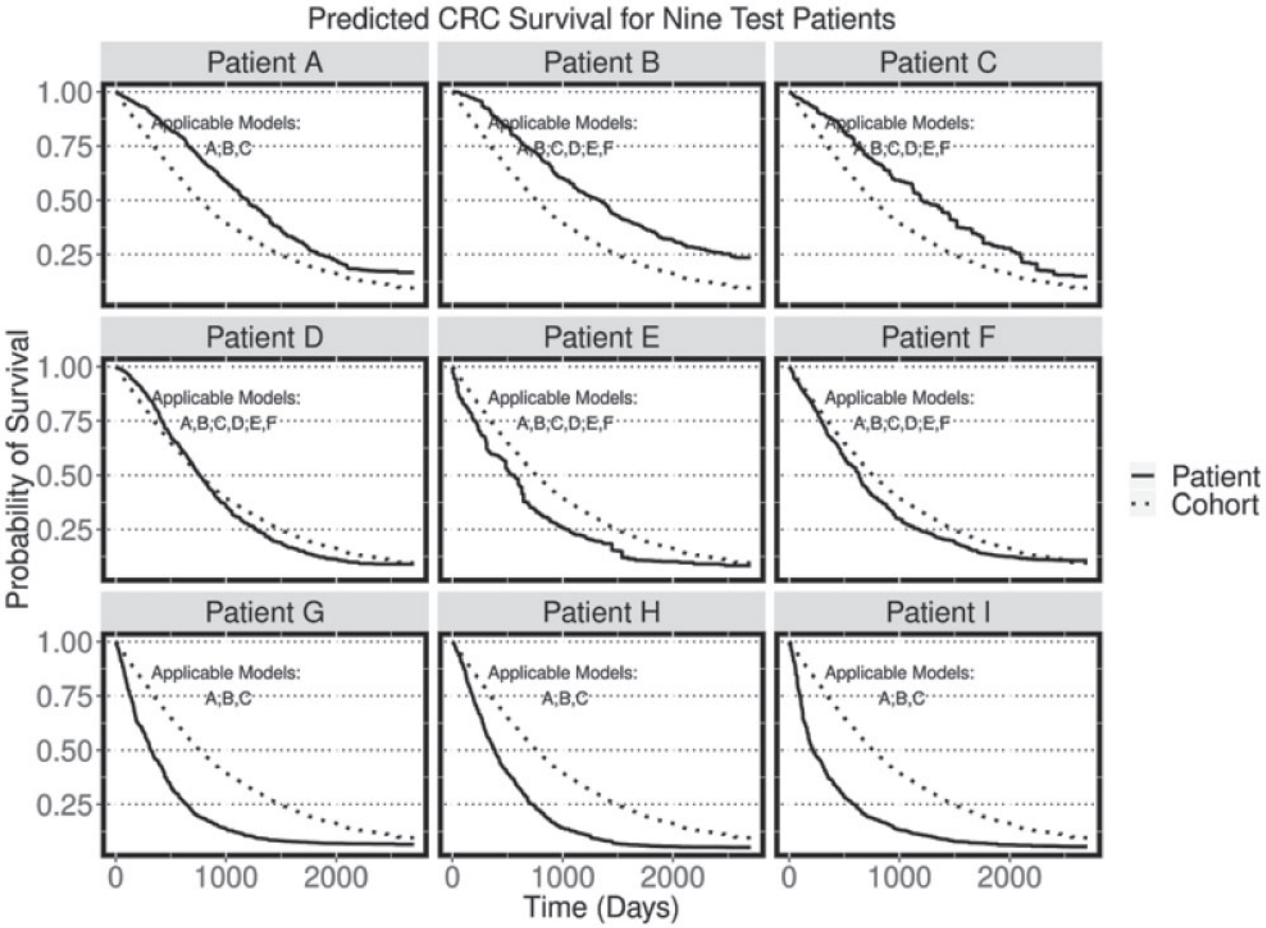


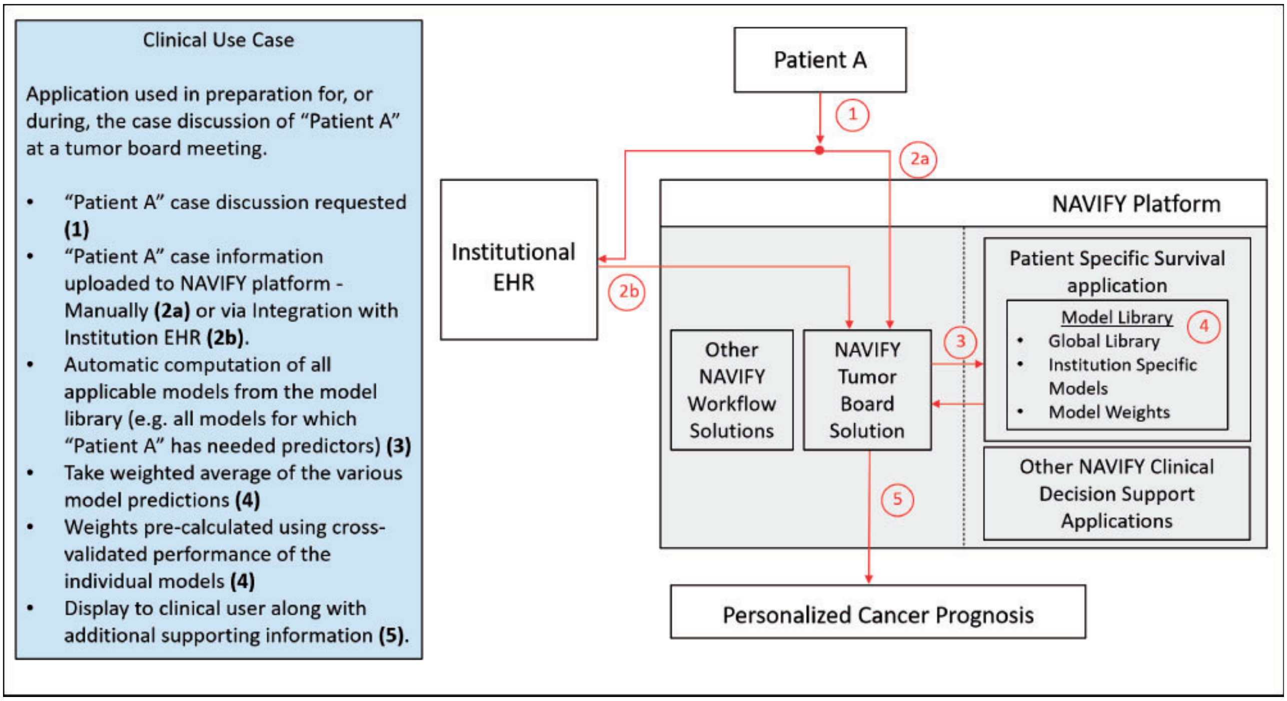
Figure 5. Patient-specific Kaplan-Meier curves for 9 selected colorectal cancer (CRC) test patients using meta-model–predicted survival probabilities. We selected from the test partition of the CRC data 3 patients with substantially favorable (top row), 3 patients with substantially unfavorable (bottom row), and 3 patients with generally typical (middle row) predicted survival. The solid lines represent the patient's predicted survival and the dashed lines show survival for the cohort as a whole. Applicable models represent the underlying individual CRC models for which the patient had the necessary predictor data and which were included in the prediction.

this may be partly due to the fact that underlying models using differing predictors may provide predictions with at least partially uncorrelated errors. Thus, averaging the "noisy" predictions produced by the underlying models may serve to reduce the overall noise (ie, overall error). While we selected our predictor set groups manually, in large part based on patterns of data availability, it may be a useful subject of future work to explore characteristics of ideal predictor sets. For example, should correlated predictors be preferentially included in the same or in different predictor sets?

To be sure, our approach is not the first attempt to build patient-specific Kaplan-Meier curves.[30,31] However, most if not all prior attempts to develop patient-specific survival curves have been based on linear models, in contrast to our primary approach. The primary novelty of this article is the use of the meta-model; however, the concept of using patient-specific survival predictions for clinical decision support may also prove to be a useful, practical application of this work.

In addition to the need for future work to further generalize the meta-model approach, the specific application to patient-specific survival prediction is subject to limitations. A key consideration is that in some cases, the algorithms may be providing information that the clinician already knew or suspected; for example, clinicians can of course in some cases use judgement to identify patients who appear sicker and likely have a worse prognosis. While formally testing the clinical value of these algorithms may be a subject of a future study, given the multitude of predictors that went into the algorithms, we hypothesize that it would be difficult for a clinician to manually integrate the value of the many predictors included in our models. Indeed, studies have shown that the human brain is unable to simultaneously integrate a large number of data elements.[2,32] We are considering performing user simulation studies to evaluate how our algorithms perform in comparison to manual clinician intuition.

We are considering several extensions to the patient-specific survival models. In particular, we may develop models for other tumor types and that incorporate additional predictors, including additional biomarkers, comorbidities, tumor genomics, patient socioeconomic factors, care delivery characteristics, and potentially even features extracted from radiologic and whole slide images. Moreover, in addition to providing prognostic predictions, we hypothesize that our approach will be applicable to patient-specific treatment optimization and prescriptive decision support. For example, we plan to explore whether we can update our models to in-



Clinical Use Case

Application used in preparation for, or during, the case discussion of "Patient A" at a tumor board meeting.

- "Patient A" case discussion requested **(1)**
- "Patient A" case information uploaded to NAVIFY platform - Manually **(2a)** or via Integration with Institution EHR **(2b)**.
- Automatic computation of all applicable models from the model library (e.g. all models for which "Patient A" has needed predictors) **(3)**
- Take weighted average of the various model predictions **(4)**
- Weights pre-calculated using cross-validated performance of the individual models **(4)**
- Display to clinical user along with additional supporting information **(5)**.

Patient A
1
2a
Institutional EHR
2b
NAVIFY Platform
Other NAVIFY Workflow Solutions
NAVIFY Tumor Board Solution
3
Patient Specific Survival application
Model Library
4
Global Library
Institution Specific Models
Model Weights
Other NAVIFY Clinical Decision Support Applications
5
Personalized Cancer Prognosis

Figure 6. A proposed schematic of model application infrastructure. Shown is a schematic of an potential infrastructure and workflow within which to implement the meta-model. This is based on the Roche Navify Tumor Board Solution.

clude as predictors the treatment the patient received and then apply counterfactual learning to predict response to therapy.

As we plan our implementation strategy, we will need to carefully evaluate how clinicians and their patients would consider insights provided by our models, and whether such knowledge would have unintended consequences. For example, how would a clinician communicate a personalized life expectancy to a patient, and how will the patient feel about this deeper understanding of their own mortality? Could this information inadvertently bias clinicians when they take treatment decisions to be more or less aggressive than they otherwise might? Who should oversee the appropriateness of CDS tools for clinical use and monitor for unforeseen outcomes? Addressing these questions may ultimately prove more challenging than the technical aspects of this clinical decision support.[33]

CONCLUSION

The "meta-model" approach we developed and demonstrated in this article offers a strategy to develop clinical predictive models that can accommodate interpatient heterogeneity in data availability and "missing data." We further demonstrate the value of random forest–based survival models in predicting patient-specific oncology survival. We expect that the proofs of concept we develop here will provide a foundation for novel types of clinical decision support to enable clinicians to make more personalized patient care decisions.

FUNDING

This work was supported using internal Roche Diagnostics funding.

AUTHOR CONTRIBUTIONS

JMB, DH, and MP contributed to general project framework. JMB with input from DH and MP contributed to development of meta-model approach. JMB contributed to model training and validation. All authors contributed to development of model application and article drafting and/or revision.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

All authors are except JMB are employed by Roche Diagnostics. JMB is a consultant for Roche Diagnostics. MP is also a Director of Open Medical Holdings Ltd, a UK-based digital health company.

REFERENCES

1. Baron JM, Dighe AS. The role of informatics and decision support in utilization management. *Clin Chim Acta* 2014; 427: 196–201.
2. Baron JM, Dighe AS, Arnaout R, *et al.* The 2013 symposium on pathology data integration and clinical decision support and the current state of field. *J Pathol Inform* 2014; 5 (1): 2.
3. Baron JM, Kurant DE, Dighe AS. Machine learning and other emerging decision support tools. *Clin Lab Med* 2019; 39 (2): 319–31.
4. Baron JM, Mermel CH, Lewandrowski KB, Dighe AS. Detection of preanalytic laboratory testing errors using a statistically guided protocol. *Am J Clin Pathol* 2012; 138 (3): 406–13.
5. Kohane IS. Health care policy. Ten things we have to do to achieve precision medicine. *Science* 2015; 349 (6243): 37–8.
6. Louis DN, Gerber GK, Baron JM, *et al.* Computational pathology: an emerging definition. *Arch Pathol Lab Med* 2014; 138 (9): 1133–8.

7. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. *Am J Clin Pathol* 2016; 145 (6): 778–88.
8. Rosenbaum MW, Baron JM. Using machine learning-based multianalyte delta checks to detect wrong blood in tube errors. *Am J Clin Pathol* 2018; 150 (6): 555–66.
9. Tomasev N, Glorot X, Rae JW, *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572 (7767): 116–9.
10. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. *Sci Transl Med* 2012; 4 (158): 158rv11.
11. Luo Y, Szolovits P, Dighe AS, Baron JM. 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. *J Am Med Inform Assoc* 2018; 25 (6): 645–53.
12. Waljee AK, Mukherjee A, Singal AG, *et al.* Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open* 2013; 3 (8): e002847.
13. Weber GM, Adams WG, Bernstam EV, *et al.* Biases introduced by filtering electronic health records for patients with "complete data." *J Am Med Inform Assoc* 2017; 24 (6): 1134–41.
14. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform* 2018; 6 (1): e11.
15. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007; 61 (1): 79–90.
16. Qi L, Wang YF, He Y. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Stat Med* 2010; 29 (25): 2592–604.
17. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; 18 (6): 681–94.
18. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; 20 (1): 40–9.
19. Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1975-2016). National Cancer Institute, Division of Cancer Control and Population Statistics, Surveillance Research Program. https://seer.cancer.gov/archive/csr/1975_2016/ Accessed October 25, 2020.
20. Zhu B, Wu X, Wu B, Pei D, Zhang L, Wei L. The relationship between diabetes and colorectal cancer prognosis: A meta-analysis based on the cohort studies. *PLoS One* 2017; 12 (4): e0176068.
21. Ma X, Long L, Moon S, Adamson BJS, Baxi SS. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. *medRxiv:* 20037143; 2020.
22. Therneau TM. A package for survival analysis in S. 2015. https://CRAN.R-project.org/package=survival Accessed October 25, 2020.
23. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. *Biostatistics* 2014; 15 (4): 757–73.
24. Ishwaran H, Kogalur UB. Consistency of random survival forests. *Stat Probabil Lett* 2010; 80 (13–14): 1056–64.
25. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; 2 (3): 841–60.
26. Ishwaran H, Lu M. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat Med* 2019; 38 (4): 558–82.
27. Breiman L. Stacked regressions. *Mach Learn* 1996; 24 (1): 49–64.
28. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; 56 (2): 337–44.
29. Heagerty PJ. Package survivalROC. 2015. https://cran.r-project.org/web/packages/survivalROC/survivalROC.pdf Accessed October 25, 2020.
30. Yu C-N, Greiner R, Lin HC, Baracos V. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In: NIPS'11: proceedings of the 24th International Conference on Neural Information Processing Systems; 2011: 1845–53.
31. Patient-Specific Survival Prediction (PSSP). http://pssp.srv.ualberta.ca/ Accessed October 25, 2020.
32. Hofman MA. Evolution of the human brain: when bigger is better. *Front Neuroanat* 2014; 8: 15.
33. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; 17 (1): 195.