**Author’s response to reviews**

**Title:** Development, Implementation, and Prospective Validation of a Model to Predict 60-day End-of-Life in Hospitalized Adults upon Admission at Three Sites

**Authors:**

Vincent J Major (vincent.major@nyulangone.org)

Yindalon Aphinyanaphongs (yin.a@nyulangone.org)

**Version:** 1  **Date:** 11 Jul 2020

**Author’s response to reviews:**

Development, Implementation, and Prospective Validation of a Model to Predict 60-day End-of-Life in Hospitalized Adults upon Admission at Three Sites

**Authors:**

Vincent J Major, ME, MPhil

Yindalon Aphinyanaphongs, MD PhD

**General response:**

We thank the reviewers and editor for their time and consideration of our manuscript. We greatly appreciate your suggestions to clarify several key points within the method section. Our intent for this work was to focus on the application of prospectively applying a single model to three sites (including several generalization challenges) and we are happy that our revisions preserve that focus but with a stronger, more clear methods section.

(See attached pdf of responses and manuscript pdf for highlighted changes.)

**Reviewer 1:**

This paper presented an interesting study of developing prediction model of 60-day mortality in hospitalized adults using random forest. The overall study design is well-planned, and more importantly, the application itself is clinical meaningful and can be implemented in the future.

**Author Response:**
We thank the reviewer for their time and appreciate their kind words.

Review Comment:

However, there are a few issues need to be addressed before considering publication:

1. In Methods section, the author stated that "patients with a lack of data were omitted", is there any particular reason for the exclusion? The amount of patients excluded (19.6%) was fairly large, and may affect the model performance. The authors may consider investigating the missing values in more details, for example, whether they are missing at random or not? Does the missing values provide any clinical insights? Maybe they represent some missed opportunities in predicting and preventing death? The authors could also try imputation techniques on these data.

Author Response:

A great point! We included these numbers to be as transparent as possible but we had best add some clarifying comments for the reader, thanks for the suggestion.

These admissions are omitted based on if they have any features in the prior year to form a prediction. The set of training data variables extracted from prior encounters, diagnosis, procedure, medication codes or lab results are counted and only those that exist more than 100 times are kept (0.1% of admissions, rounded up to 100). (More about this pruning later.) For each patient, their data in the year prior to each admission are assembled and if any admission has zero data (demographics are not included in this count) they are omitted. We have double checked the data available for this 19.6% of patients and can confirm they only have demographics data.

And you are correct, this data is not missing at random, it represents a notable subpopulation of patients with almost zero data. We suspect that these are patients that have: 1) never have been (or been seldom) seen at our institution, or 2) typically healthy patients with no recent data. Imputation in this group would be infeasible as we have no data to help guide a multiple imputation method, for example.

Although it is beyond the scope of this manuscript, we have ongoing work in this area to overcome this limitation that 20% of admissions have no prediction made. Our work delays the timing of prediction to add new data collected from hospital day 1. In some preliminary testing we see an almost complete coverage of predictions (99.9%) and have been able to see new phenotypes, e.g. patients presenting with sepsis secondary to pneumonia, or post mechanical fall, even when the patient is new to our system.
As a result of this and other suggestions we have simplified this statement and moved the 19.6% number to another section that can explain in further detail why these patients are omitted.

Changes

Methods &gt; Setting and Data &gt; Study Cohort

Removed: but patients with a lack of data were omitted (19.6% of the training cohort).

Methods &gt; Model Development

… Some training cohort patients (19.6%) are left with only demographic features and are removed from model training. No data imputation is performed….

Review Comment:

2. The author mentioned the random forest incorporated 9614 variables in the abstract, but did not mention how these variables were selected in the Methods section, and only a few were presented in Table 1 (I was also wondering how they authors choose which predictors/comorbidities can be listed in the table?) In addition, with that many variables, it is destined to introduce bias, have the authors considered filter or preselect a set of variables?

Author Response:

Thank you for your comment. Only a very small number of features could be shown in a table, we manually selected what we believed were the most of interest, the count of several types of code and types of encounter in the most recent time slice (1-30) and the entire prior year (1-365). We believe these numbers are the most intuitive features for readers to compare against their own experience.

In regards to the number of features used in the model, we agree we should add some clarifying comments in the manuscript. You are correct that with this many features, bias will be inevitably present in the predictions. However, we are attempting to balance bias and variance with respect to predictive performance. As alluded to in the prior response, we do perform some feature pruning before model training. In addition, the retrospective and prospective validations are the ultimate guard against overfitting and these results suggest that the model has generalized according to the model selection methodology. I would also add that a primary benefit of random forest as an algorithm is the built-in feature selection with the use of multiple decision trees and the random sampling of candidate features at each node of each tree. This characteristic is an additional safeguard against bias.

In terms of feature pruning, our methodology is checking the number of times each feature is measured in both the death and survival groups and we prune all features that are only in one
outcome group, or with a total count less than 100 (0.1% of training set admissions, ~72 which is rounded up to 100). This strategy is an attempt to prevent some features, by pure chance, only occurring in patients that die, which would bias the model to learn that feature leads to death. The count of 100 is to minimize cases where, again by chance, 9/10 die (mitigated by limiting how small the sample size can be). For reference, the raw data contains 87,226 candidate predictors and after the pruning procedure only 9,614 (11%) remain. We have added a paragraph in the Methods, Model Development section to clarify this approach as both reviewers noted similar concerns.

Changes

Methods &gt; Model Development

Many thousands of candidate predictors are expected where only a small fraction exist for a typical patient. With this many predictors, overfitting to spurious associations and small samples is difficult to avoid. Predictors are pruned by requiring at least once occurrence in both outcome groups (i.e. survival and death) and a total count exceeding 100. This leaves 9,614 features (11% of 87,226) for modeling. Some training cohort patients (19.6%) are left with only demographic features and are removed from model training. No data imputation is performed.

Review Comment:

3. The authors only used random forest as their modelling approach, have the authors tried other algorithms? For example decision tree. A common practice would be test different models and identify the best one based on both performance and interpretability.

Author Response:

You’re very right. This type of empirical testing was performed early on in our project. We compared logistic regression, gradient boosting (XgBoost) and random forest models and found random forest was consistently the best performing model within the five-fold cross validation. This is also consistent with related works (below) that we have added into the manuscript in a revised sentence.

We chose early-on to focus this manuscript on the implementation and prospective validation. It is possible to write a complete paper focused just on model selection and felt that this retrospective comparison was beyond the objective of this manuscript.


Changes

Methods &gt; Model Development

Many of these features are sparse (e.g. specific ICD-10 codes), others are complete but highly nonlinear (e.g. count of procedure codes) which complicate modeling. An algorithm is needed that can learn which features—and which values within those features—are prognostic. A random forest classifier, a type of tree-based algorithm, was employed for its consistent performance on a variety of datasets (Olson) and similar mortality work (Makar, Parikh). ...

Review Comment:

4. The authors described a different performance when applying the model to Brooklyn population, I would love to read more discussion on this as to why and what may cause the observed difference, this may provide important insights on the generalizability and representativeness of the study results.

Author Response:

You are absolutely right, the Brooklyn performance discrepancy is a particularly interesting challenge. We have pointed the reader several times to the Supplemental Online Content where there is some extra investigation. We debated including more of this analysis in the manuscript but were unfortunately restricted by word limits.

In terms of discussion, our opinion is that the underrepresentation of Brooklyn patients effectively dilutes the Brooklyn sample observed during training. We know there are structural differences in the cohorts (eTable 1) which fit with our understanding of the very different patient mix seen in Brooklyn associated with demographic and socio-economic factors. Combining these two aspects, the model’s ability to generalize to Brooklyn suffers.

We’ve added some clarifying comments in the Discussion but we are hesitant to further speculate without more concrete data.

Changes

Discussion &gt; Generalization to Brooklyn cohort
An evolving patient population is common in many applications and creates a practical challenge for prospective validation. In this case the mechanism of change is apparent: a new hospital was brought into the system that treats a new population with varying comorbidities and social determinants of health. Although this cause is obvious, the consequences are not. Along with an increase in proportion of patients observed at the Brooklyn hospital (38.0% in testing up from 9.5%; Table 1), there are corresponding structural differences in age, race, sex, outcome, and comorbidities between Manhattan and Brooklyn sites (eTable 1). These differences, indicative of a larger disparity between sites, further complicate generalization. Despite this, a model trained with only 10% of cases being from a new hospital can adapt to be performant and safely applied in a shifted patient mix.

The underrepresentation of Brooklyn patients during training does affect model performance at the Brooklyn hospital and their representation in the identified high-risk group (eFigures 5 and 6). A larger sample of Brooklyn patients for model training may improve the model’s ability to learn the new site and improve performance. ...

Review Comment:

5. The authors may also consider explain how and why the model works in clinical practice, for example, the authors could try some explainable techniques and models in addition to random forest. An explainable component would improve the feasibility of the application as many clinicians would like to know how the model works before using it.

Author Response:

Thank you for your thoughts. Explainability has a frequently debated issue during this project and ongoing work. The intention of this work was to start identifying patients prospectively for review by physicians. In our preliminary work, we have found that many of these patients have very complex disease that, in many ways, is obvious to an experienced physician. What is not obvious is if they are at high vs. moderate risk of dying in 60 days. While the model can perform this kind of discrimination, we have found that the most influential features (using methods like variable importance, ablation or SHAP values) often do not convey these differences. Instead, the explanation adds complexity and cognitive load that is easy to find fault with, even if the prediction is correct. For example, the most influential features of this model (eTable 2) are not terribly intuitive, e.g. max # of diagnosis codes.

For our intended application of prompting supportive care and end-of-life planning, we have decided, as part of a larger collaborative effort, to focus on communicating the predictions and measuring user behaviour towards a goal of consistency. But, we will absolutely continue to explore explainable techniques.
Review Comment:

Other minor issues:

1. Table 1 has too many comparisons, the p-values reported should be adjusted or corrected accordingly.

Author Response:

Thank you, this table grew and we overlooked that we should make adjustments. We have updated the Table 1 (and eTable 1) with Bonferroni corrected p-values using 36 (and 29) trials. The large sample size continues to produce very small p-values that are well within the bounds of the adjusted significance threshold of 0.05/36=0.0014 (or 0.0017).

Changes

Updated Table 1 and eTable 1

Review Comment:

2. The authors may want to work on the some language corrections, for example, the first sentence in methods section of abstract (A prognostic study of .......within 60 days of admission) is hard to read.

Author Response:

We thank the reviewer for their suggestions. We have made numerous changes through the manuscript to improve its readability. Many are which are highlighted in the text but not listed here for brevity.

Reviewer 2:

This is a very interesting study that predict the mortality in 60 days after admission using multi-site EHR data. In addition to evaluation within the studying cohort, the authors further conducted prospective validation of their machine learning model. Generalizability of the ML model across different sets were assessed. The impacts of demographics were also investigated. Overall, I believe this study has a good quality. Although this manuscript was in generally written well. The organization of the manuscript can be improved for an informatics journal.
Author Response:

We thank the reviewer for their time and consideration.

Review Comment:

Here are my major comments:

1. I feel like this paper is organized as a medical journal rather than a medical informatics journal, where the methodologies should be highlighted. Important descriptions on methods are not available in the manuscript. The authors should consider moving major methods details back to the manuscript.

Author Response:

Thank you for your suggestions. We admit we made some difficult decisions to satisfy the word limit but have made numerous changes to the manuscript to better clarify key factors of the methodology. In particular, we made sure that every section of the eMethods was mentioned in the main methods section in what we believe is sufficient detail for most readers.

In terms of organization, sections such as the setting of this study are particularly important for context later in the manuscript as we discuss the generalization across site. We concede that these components give the manuscript the feel of a medical journal but we would not want to remove anything or reorganize in a way that endangers the context later on. We have revised our naming of section headings to better align with the Medical Informatics audience.

Changes

Methods &gt; Data &gt; Study Setting

Methods &gt; Data &gt; Patient Population

Methods &gt; Data &gt; Mortality Outcomes

Methods &gt; Data &gt; Feature Construction

Patient demographics and discrete data describing prior encounters are collected along with several categories of coded data used in related works (7,10,11), namely: .... Each data type, except demographics, are dated in a patient’s history and can occur many times.

Features are constructed from these data similarly to that of Avati et al. (7) and other related works (7,10,11). Specifically, each patient’s history is segmented into four time slices with
boundaries at 30, 90, and 180 days preceding admission (7), excluding all data collected more than a year prior. Each data category (e.g. ICD-10 diagnoses) is aggregated in each slice into:

Count of each unique code,

Count of unique codes and total code count across days, and

Mean, variance, minimum, maximum, and range of the daily number of codes.

While a typical patient may have fewer than a dozen unique ICD-10 codes, all patients have aggregate values, e.g. total codes during slice and max number of daily codes. With these features, a model may learn differences in specific disease types as well as disease burden and utilization with these features.

Methods &gt; Experimental Design

A retrospective cohort-study experimental design that ‘enrolls’ each admission is employed with a temporally separated testing set. ... No other cohort selection criteria were applied during the training period—all (re)admissions were included. Patients who were readmitted within the testing period (2017) following their ‘enrollment’ during the training period (2015/2016) are excluded to ensure no individual patient is present in both groups.

Methods &gt; Model Development

Many thousands of candidate predictors are expected where only a small fraction exist for a typical patient. With this many predictors, overfitting to spurious associations and small samples is difficult to avoid. Predictors are pruned by requiring at least once occurrence in both outcome groups (i.e. survival and death) and a total count exceeding 100. This leaves 9,614 features (11% of 87,226) for modeling. Some training cohort patients (19.6%) are left with only demographic features and are removed from model training. No data imputation is performed.

Many of these features are sparse (e.g. specific ICD-10 codes), others are complete but highly nonlinear (e.g. count of procedure codes) which complicate modeling. An algorithm is needed that can learn which features—and which values within those features—are prognostic. A random forest classifier, a type of tree-based algorithm, was employed for its consistent performance on a variety of datasets (15) and similar mortality work (10,16).

Methods &gt; Evaluation in the Context of Potential Demographic Bias

... model performance in different strata of sensitive demographics is investigated (eMethods, eResults, eFigures 5 and 6). One model, trained on the entire training set is applied to sub-cohorts of the testing cohort by combinations of sensitive demographics (e.g. Black women admitted in Brooklyn) and various measures of model performance are reported. This procedure is repeated for a second model where all sensitive demographics are excluded or ‘masked’ during training (eMethods and eResults).
A random forest model was selected for its robust performance in sparse, high dimensional feature spaces. To mitigate over-fitting in very specific subpopulations, predictors that did not occur more than 100 times in the training set were pruned before training leaving 9,614 predictors.

Review Comment:
2. It's not clear how the last admission (whether mortality happens within 60 days) was selected for both patient with and without mortality outcome.

Author Response:
Good question, and one that is often overlooked. We perform no such selection in our experimental design. Every admission is included and the outcome calculated however it falls, regardless of the death or survival of the patient during follow-up. If a patient that ultimately dies has several admissions during the training period at, let’s say, 180, 80, and 20 days before death, all three will be included and their outcomes will be 0, 0, and 1. The model is forced to learn what has changed about the patient to make their label flip from negative to positive.

This type of retrospective cohort study design includes many more admissions than one that, for example, randomly selects one admission per patient. The extra sample should help the model learn trajectories of patient decline. A retrospective cohort study also helps to ease transition into a subsequent prospective cohort study (as is the case in this work) as the nuances of readmissions, for example, are seen in both studies.

We have added a clarification in the Methods section.

Changes

Methods &gt; Experimental Design

A retrospective cohort-study experimental design that ‘enrolls’ each admission is employed with a temporally separated testing set. ... No other cohort selection criteria were applied during the training period—all (re)admissions were included. Patients who were readmitted within the testing period (2017) following their ‘enrollment’ during the training period (2015/2016) are excluded to ensure no individual patient is present in both groups.

Review Comment:
3. I feel that the use of random forest is not well justified. There exist many conventional machine learning algorithms (e.g. logistic regression) and deep learning algorithms (e.g. recurrent neural network) for clinical outcome prediction. As an informatics journal, the authors should provide more details on the justification of their selection of method.

Author Response:

You are correct that many algorithms have been used for various applications. The intent of this manuscript was not to compare alternative algorithms. But, early on in this project, we did perform some experiments comparing logistic regression, gradient boosting (XgBoost) and random forest models and found random forest was consistently the best performing model. We did not consider deep learning algorithms, but personal communication with authors of other related work has led us to believe that complex deep learning methods yield little gain over robust tree-based methods such as random forest and gradient boosting (Nigam Shah, 2018, personal communication). In addition, the multitude of models, especially non-linear ones, in fairly limited feature spaces will generally produce very similar performances. Moreover, any performance differences in AUROC manifest as small changes in parts of the ROC or PR curves that are not relevant to the prospective application (as the various models identify the same “high” risk patients).

Other relevant references:


Changes

Methods &gt; Model Development

Many of these features are sparse (e.g. specific ICD-10 codes), others are complete but highly nonlinear (e.g. count of procedure codes) which complicate modeling. An algorithm is needed that can learn which features—and which values within those features—are prognostic. A random forest classifier, a type of tree-based algorithm, was employed for its consistent performance on a variety of datasets (Olson) and similar mortality work (Makar, Parikh). ...
Review Comment:

4. It was not very clear to me how the experiments on generalizability of model were conducted. Did the authors train one model using the training data from one location and apply the trained model to the testing data of another location? Or did the authors train one model using training data from all locations and apply the trained model to the testing data of another location?

Author Response:

Your latter alternative is correct. The final model (that was prospectively tested) was tested within subgroups by site (and other demographics). No site-specific models were developed as our intention was to have one universal model deployed at each site. (A model ‘masked’ of race and ethnicity was developed for comparison but its results are only within the Supplement.)

Unfortunately due to word limits, much of the detail around these experiments are within the Supplemental Online Content. But, we have added some clarification of the experiments in the main text.

Changes

Methods &gt; Evaluation in the Context of Potential Demographic Bias

... model performance in different strata of sensitive demographics is investigated (eMethods, eResults, eFigures 5 and 6). One model, trained on the entire training set is applied to subcohorts of the testing cohort by combinations of sensitive demographics (e.g. Black women admitted in Brooklyn) and various measures of model performance are reported. This procedure is repeated for a second model where all sensitive demographics are excluded or ‘masked’ during training (eMethods and eResults).

Review Comment:

5. It would be interesting to see stratified analyses on data category, such as diagnosis codes, medication codes, etc. As

Author Response:

Unfortunately, it appears part of your comment was cut off. What we believe you are asking about is an analysis of the additional gain of each data category. If we have misunderstood, we apologize in advance.
This is something that we have considered in the past and there are two main analyses we have discussed. The first approach is considering a baseline model of demographics, what is the ‘lift’ of each data category and then the extra lift of adding another? The second is learning a total model and ablating each data category to assess the model’s reliance on that data. We believe the latter gets closer to the core question of: “what is the model learning?” as we are identifying what features the model cannot do without.

We completed a brief analysis for you removing each of the five data types (encounters, diagnoses, procedures, medications, and lab results) from the testing set without retraining the model. We found that the largest drop in AUROC and AUPRC performance occurred when removing encounter information (1.2 and 4.7%). Removing diagnoses (1.4 and 4.0%) and lab results (0.8 and 3.5%) also caused relative large decreases. Removing procedures caused a small decrease (0.1 and 0.8%) and, finally, removing medications caused very little change (0.0 and 0.2%). Overall, medications and procedures are the least beneficial data categories when ablated from the testing cohort data.

We consider these experiments to be focused on model development and keeping with our intent to focus the core manuscript to the implementation and prospective validation we have added these results as a table to the Supplemental Online Content, eTable 3, reporting median [95% CI] AUROC and AUPRC for each ablation condition.

Changes

Results &gt; Retrospective Modeling &gt; Performance Within the Testing Cohort

Each category of data contributes to overall redundancy where removing any one has little to no effect on overall performance (eTable 3).

eTable 3. Model performance after ablating one category of data.

Performance of the final model when applied to five testing sets each with a different category of data removed.

Cohort

Ablation

Measure

AUROC

AUPRC

Testing

None (Table 2)
Median
[95% CI]
87.2
[86.1, 88.2]
28.0
[25.0, 31.0]

Encounters
Median
[95% CI]
86.0
[85.6, 86.5]
23.3
[22.1, 24.4]

Diagnoses
Median
[95% CI]
85.8
[85.3, 86.3]
24.0
[22.9, 25.3]

Procedures
Median
[95% CI]
87.1
[86.6, 87.6]
27.2
[25.9, 28.5]

Medications
Median
[95% CI]
87.2
[86.7, 87.7]

27.8
[26.6, 29.2]

Lab Results
Median
[95% CI]
86.4
[85.9, 86.8]

24.5
[23.2, 25.7]