

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

# **BMJ Open**

#### Garbage Input, Variable Output: Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU Mortality

Journal:	BMJ Open
Manuscript ID	bmjopen-2020-041421
Article Type:	Original research
Date Submitted by the Author:	10-Jun-2020
Complete List of Authors:	Iwashyna, Theodore; University of Michigan, Internal Medicine; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Ma, Cheng; University of Michigan, Statistics Wang, Xiao Qing; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Seelye, Sarah; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Zhu, Ji; University of Michigan, Department of Statistics Waljee, Akbar; University of Michigan, Gastroenterology; VA Center for Clinical Management Research, Ann Arbor VA Medical Center
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, STATISTICS & RESEARCH METHODS, Adult intensive & critical care < INTENSIVE & CRITICAL CARE





I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

reliez oni

**Title:** Garbage Input, Variable Output: Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU Mortality

#### Authors:

Theodore J. Iwashyna\*, M.D., PhD (1,2,3)

Cheng Ma\*, B.S. (4)

11 Xiao Qing Wang, MPH (1)

Sarah Seelye, Ph.D. (1)

Ji Zhu, Ph.D. (4)

Akbar K. Waljee, M.D., M.Sc. (1,2,3)

<u>\*considered co-first authors.</u>

- 1) VA Center for Clinical Management Research, VA Ann Arbor Health Care System, Ann Arbor, MI, USA.
- 2) Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA.
- 3) Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), Ann Arbor, MI, USA
- 4) Department of Statistics, University of Michigan, Ann Arbor, MI, USA.

Address correspondence to: Akbar K. Waljee, M.D. M.Sc., Associate Professor of Medicine, Division of Gastroenterology,
 Department of Internal Medicine, Michigan Medicine, Ann Arbor Veterans Affairs Medical Center, 2215 Fuller Road, 111D,
 Ann Arbor, Michigan 48105

Phone: 734-845-5865; Fax: 734-845-3091; E-mail: <u>awaljee@med.umich.edu</u>

Keywords: missing data, risk prediction, machine learning, electronic health record data, random forests

Word count: 3,085

#### BMJ Open

#### ABSTRACT

2 3	<b>Objective:</b> There has been a proliferation of approaches to statistical methods and missing data imputation as electronic
4 5 6	health records become more plentiful. The relative performance on real-world problems is unclear.
7 8	Materials and Methods: Using 355,823 ICU hospitalizations at over 100 hospitals in the nationwide VA healthcare
9 10 11	system (2014-2017), we systematically varied 3 approaches: how we extracted and cleaned physiologic variables; how
12 13	we handled missing data (using mean value imputation, random forest, extremely randomized tress (extra-trees
14 15	regression), ridge regression, normal value imputation, and case-wise deletion); and how we computed risk (using
16 17 18	logistic regression, random forest, and neural networks). We applied these approaches in a 70% development sample
19 20	and tested the results in an independent 30% testing sample. Area under the ROC Curve (AUROC) was used to quantify
21 22	model discrimination.
23 24 25	<b>Results:</b> In 355,823 ICU stays, there were 34,867 deaths (9.8%) within 30 days of admission. The highest AUROC's
26 27	obtained for each primary classification method were very similar: 0.83 (95% CI [0.83-0.83]) to 0.85 (95% CI 0.840.85).
28 29 20	Likewise, there was relatively little variation within classification method by the missing value imputation method
30 31 32	used—except when case-wise deletion was used for missing data.
33 34	Discussion: Variation in discrimination was seen as a function of data cleanliness, with logistic regression suffering the
35 36 37	most loss of discrimination in the least clean data. Losses in discrimination were not present in random forest and neural
38 39	networks even in naively extracted data.
40 41 42	Conclusion: Data from a large nationwide health system revealed interactions between missing data imputation
42 43 44	techniques, data cleanliness, and classification methods for predicting 30-day mortality.
45 46	
47 48 49	
50 51	
52 53	
54 55 56	
50 57 58	2
59 60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study focuses on a large, real world data set consisting of 355,823 ICU stays at over 100 different facilities. •
- Multiple methods of model fitting and missing data imputation were implemented in standardized ways that . reflect common practice.
- The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility, and we encourage validation on other data sets.
- Due to high dimensionality of method combinations, this study only considered one outcome, and only • considered one standardized and decided upon a priori approach within each dataset / categorization model / missingness imputation triad.

**BMJ** Open

#### INTRODUCTION

Risk adjustment plays an increasingly central role in the organization, care of, and science about critically ill patients[1, 2]. Statistical adjustment is essential for many performance measurement as well as pay-for-performance and shared savings systems, from US News and World Report to Medicare and Medicaid. It is used to stratify the care of patients for treatments and track quality improvement efforts over time[3]. It is routinely measured, even in clinical trials, to assess confounder balance between arms and may form part of RCT enrollment or drug approval criteria[4].

As a result, there has been a proliferation of risk scores both for the common task of short-term mortality prediction and for assorted more specialized tasks. Many statistical tools have been promoted. Rules of thumb have developed and existed long enough to be critiqued[5-9]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines offer standardization of reporting[10]. Textbooks have emerged[11]. Yet questions remain on fundamental pragmatic issues: How clean does the data have to be to prevent the so-called "garbage in, garbage out (GIGO)" phenomenon? How sensitive are methods to missing data and how should it be handled? Do these analysis decisions interact?

To address such questions, we compared the performance of an array of methods on a single standardized common problem—the prediction of 30-day mortality from day 1 laboratory results among patients admitted to the Intensive Care Unit (ICU) at any hospital in the nationwide Veterans Health Administration system[12-14]. Using exactly the same set of real ICU admissions, we systematically varied three parameters: the approach used to extract and clean physiologic variables from the electronic health record; the approach used to handle missing data; and the approach used to compute the risk. We systematically applied these approaches in a 70% development sample and tested the results in an independent 30% testing sample, to provide real world comparisons to inform future pragmatic implementation of risk scores.

#### METHODS

### Cohort

Data were drawn from the Veterans Affairs Patient Database (VAPD 2014-2017), which contains daily patient physiology
for acute hospitalizations between January 1, 2014 and December 31, 2017. The VAPD 2014-2017 includes patient
demographics, laboratory results, and diagnoses that are commonly used to predict 30-day mortality from the day of
admission. Here, we included data from all ICU hospitalizations on day 1 of each hospitalization. Full details of the VAPD
2014-2017 have been published elsewhere[15].
The development of this data was reviewed and approved by the VA Ann Arbor Healthcare System's Institutional Review
Board.
Four versions of the dataset were created for each hospitalization on admission: A) raw lab values extracted using only
lab test names, B) raw lab values extracted using only Logical Observation Identifiers Names and Codes (LOINC), C)

cleaned lab values extracted using both LOINC[16, 17] and searched text lab test names, and D) cleaned lab values

converted to Acute Physiology And Chronic Health Evaluation (APACHE) points, extracted using both LOINC and lab test

names.

#### No Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

#### **Predictor Variables**

In our primary analyses, we adjust for 10 laboratory values that were collected within one day of hospital admission. Further patient-level adjustments included demographic characteristics (gender, age, race, and Hispanic ethnicity), 30 comorbidities, and 38 primary diagnoses. The individual comorbidities used in models are defined by methods described

#### **BMJ** Open

in van Walraven's implementation of the Elixhauser comorbidity score[18]. We adjust for 38 primary diagnoses drawn from the Healthcare Cost and Utilization (HCUP) Clinical Classification Software (CCS)[19], which consist of the top 20 most frequent single-level CCS diagnoses and 18 level-one multi-level categories of diagnoses (Appendix A.) In secondary analyses, to emphasize the role of data cleanliness, we estimate risk using only the laboratory values since the nonlaboratory values do not vary in data cleanliness and curation. 

#### **Outcome Variable: 30-day mortality**

Our primary outcome variable is 30-day all-cause mortality, defined as death within 30 days of the admission date for the index hospitalization. Mortality is evaluated using the highly reliable Veterans Administration beneficiary death files which aggregate from several sources[12, 20, 21].

#### **Statistical Analysis and Model Development**

Random Forests is an ensemble machine learning method that aggregates the results of multiple decision trees fit on bootstrap samples of the original data[22, 23]. For each decision tree, the original data are bootstrapped to create a new dataset of the same size and the tree is fit to the new data. Instead of considering all predictors to determine the splitting criterion at a node, the split variable is chosen from a random subset of variables in order to reduce the correlation between different trees. Many such trees are grown, creating a 'forest'. Each observation is classified by each tree, and the majority classification over all trees is the predicted class. The ability of random forests to learn nonlinear and complex functions contributes to its predictive performance.

The neural network[24] can "learn" to classify samples without manual designed task-specific rules. The algorithm applies different weights to predictors and uses these transformations in subsequent "layers" of the neural net, culminating in the output layer with predictions. We applied the random forest and the neural network on our task. A traditional logistic regression model was also performed and compared.

Statistical analyses were performed with Python and the scikit-learn package[25].

#### **Training and Testing Sets**

The dataset was randomly split into a 70% training set and a 30% testing set. The same split was used for all classification

methods. This process was replicated five times (five different training sets and corresponding testing set were generated),

and each time the models were fit on the training set and used to predict the 30-day mortality of the testing set.

#### Missing Data and Imputation

We imputed the missing values before training and testing the models, comparing:

- "Mean Value": the mean value of each variable in the training set was used to replace missing values[26].
- "Random Forest": use random forest to impute missing values (missForest)[27].
- "Extremely Randomized Trees (Extra-Trees Regression)": this method is similar to random forest but is faster[28, 29].
- "Ridge Regression": use Bayesian Ridge regression to impute missing values[30].
- "Normal Value"[31]: use normal values to impute missing values—this is common in clinical prediction contexts in which it is assumed that clinicians order tests they fear are not normal, and therefore the absence of such a test is a sign that the clinician reviewed other aspects of the patient's case and judged the odds of physiologic abnormality so low that testing was not indicated.
- "No Missing": case-wise deletion[32].

#### Variable Importance and Partial Dependence Plots

Predictor variable importance is evaluated for random forests[33]. When classifying a sample using a decision tree, a predictor is used at each node. Predictors that appear more frequently and that reduce the misclassification more substantially are considered more important. By combining all trees in a random forest model, we assessed the variable

BMJ Open

importance of each predictor. We also plotted the Partial Dependence Plots[30] to show how the value of predictors

affects 30-day mortality. Partial dependence plots are used to visualize assess non-linearity among variables.

#### RESULTS

#### **Cohort Description**

The cohort involved 355,823 ICU hospitalizations at over 100 different hospitals, as has been described elsewhere. The

mean age of the cohort was 66.9 years, and there were 34,867 deaths within 30-days of admission, a primary outcome

event rate of 9.8% (Table 1.)

## Table 1. ICU Patient Demographics

Variables	ICU Only Cohort
Hospitalizations, N	355,823
Age, mean (SD), y	66.9 (11.6)
Male, N (%)	341,579 (96.0)
Race, N (%)	
White	256,293 (72.0)
Black or African American	73,855 (20.8)
Other	25,675 (7.2)
Hispanic, N (%)	20,532 (5.8)
30-day Mortality, N (%)	34,867 (9.8)
Length of Stay, mean (SD), days	9.5 (13.0)

Rates of data missingness for each laboratory value in each dataset are shown in Table 2.

#### Table 2. Proportion of Labs Missing

Dataset	Albumin (albval)	Bilirubin (bili)	Blood urea nitrogen (bun)	Creatnine (creat)	Glucose (glucose)	Hematocrit (hct)	Partial Presssure (pao2)	pH (pa)	Sodium (na)	White Blood Cell (wbc)
А	0.39	0.42	0.84	0.13	0.07	0.85	0.66	0.14	0.11	0.13
В	0.38	0.42	0.13	0.13	0.06	0.12	0.65	0.44	0.11	0.13
С	0.39	0.45	0.13	0.12	0.06	0.11	0.69	0.64	0.11	0.13

#### Using all Data for Model Development

Figure 1 shows the AUC scores of different classification models and imputation methods in the primary analysis. The highest AUC's obtained for each primary classification method (rows of the figure: logistic regression, random forest, or a neural network) were very similar: AUC's of 0.83 to 0.85. Likewise, there was relatively little variation within classification method by the missing value imputation method used, be it mean value imputation, random forest, extremely randomized trees (extra-trees regression), ridge regression, or normal value imputation. All models suffered dramatic losses in discrimination when case-wise deletion was used for missing data in the least clean dataset (far right columns). Full model performance for each condition can be seen in Appendix B.

Variation in discrimination was seen, however, across classification methods, as a function of data cleanliness. (Note that the analyst was blinded to which dataset was which during the analysis). In the logistic regression model developed using the least clean data (dataset A had raw lab values extracted using only lab test names), performance was always lower than the performance with the more complete and clean datasets—by AUC's of 0.05 to about 0.1, p-value < 0.05). Similarly, performance in dataset B (extracted using LOINC codes without unit standardization) was lower and more unstable for mean value imputation and ridge regression. In marked contrast, neither random forests nor neural networks showed such reduced performance when developed in less clean data—in no case did the AUC degradations exceed 0.025 despite similar optimal performance.

#### Secondary Analysis Using only Laboratory Values

The primary analysis presented above considers the real world case in which demographics, diagnoses, and laboratory values are used in combination in risk model prediction. Yet, of these, only laboratory values were subject to variation in cleanliness; therefore we conducted a secondary analysis using only laboratory values in order to bring more clearly into relief the impact of data quality. Results are shown in Figure 2.

#### **BMJ** Open

Average model performance with this much smaller group of predictors is, as expected, somewhat lower with less data optimal AUC's typically range from 0.73 to 0.78 across combinations of classification model and missing data imputation. No uniformly superior strategy is evident, save markedly lower performance of case-wise deletion in the least clean dataset (A). As before, logistic regression shows markedly reduced discrimination when developed in the least clean data set. Neural networks show consistent performance.

Also notable is the marked reduction of discrimination of random forest models and neural network models regardless of missing data imputation model in dataset D. Dataset D is the "cleanest" data, in that it has hand-curated inclusion criteria, standardization of units, but then also conversion of all values from their continuous scale to a semi-quantitative set of "points" as is done in the APACHE scoring algorithms. Attempting to work with such standardized point values as inputs consistently resulted in markedly worse discrimination in random forest models and neural network models than using other "less clean" datasets (The difference between Dataset D and other datasets is significant with a p-value < 0.05).

#### Variable Importance

The most important predictors were age and laboratory values. Age had the highest importance scores, regardless of which dataset was used, indicating that age is the most important variable when predicting 30-day mortality. The 10 laboratory values also got high importance scores. For datasets A, B, and C, they fell in the top-13 most important variables, and there were at least eight laboratory values in the top-10 most important variables. However, for dataset D, there were only 6 laboratory values in the top-10 most important variables, and the variable white blood cell score ranked the 20<sup>th</sup>. This may indicate that transforming laboratory values to APACHE scores results in the loss of information contained in the original values and negatively influence the performance of the random forest model.

#### Partial Dependence Plots

As it is hard to visualize the relationship between multiple predictors and the outcome, we created partial dependence plots to show the effect of predictors on the outcome[34]. The plots can also show whether the relationship between a specific predictor and the outcome is linear, quadratic, monotonic, or more complex. Further analysis can be done by

combining the partial dependence plots and medical knowledge. **Figure 3** and **Figure 4** are the partial dependence plots for the pH score and the PaO<sub>2</sub> score. We will take these as examples to show how the value of predictors in different datasets affects 30-day mortality. The X-axis is the value of the predictor. For each value of the predictor, the Y-axis is the averaged model output for all observations with the corresponding value of the predictor. As we know, the normal value of the pH score is 7.4, and both higher value and lower value are abnormal. Therefore, a U-shaped partial dependence plot is to be expected for datasets A, B, and C. However, only the plot for dataset C is U-shaped. It is because the dataset C is the cleanest one, and the models can learn the real effect of pH score on the 30-day mortality. Datasets A and B are not as clean as dataset C, as some other variables are presented in these datasets as pH score. Thus, it is difficult for the models to utilize the pH score variable in datasets A and B. This result indicates that cleaner variable benefits the classification models. However, not all variables have this problem. For most other variables such as the PaO<sub>2</sub> score, the plots of datasets A, B, and C have similar trends.

#### DISCUSSION

We used real data from a large nationwide health system to explore the interaction between missing data imputation techniques, data cleanliness, and classification methods for the common problem of predicting 30-day mortality in a held-out testing dataset. In brief, we found that any of several imputation techniques other than case-wise deletion performed equivalently in terms of discrimination, regardless of data cleanliness or classification method to be used. We found that logistic regression showed worse discrimination with less carefully cleaned data than did random forest or neural networks. Random forest models (and to a degree, neural networks) displayed diminished discrimination when given data that had been too highly cleaned and standardized prior to use.

#### **Relationship to Past Research**

Missing data are ubiquitous in large datasets. Even when missingness is completely at random, missing lead to significant loss in statistical power and predictive ability[32]. We have previously found that the Random Forest method

#### Page 13 of 33

#### **BMJ** Open

consistently produced the lowest imputation error compared to commonly used imputation methods[26]. Random Forest had the smallest prediction difference when 10-30% of the laboratory data was missing. Yet our present analysis of real data shows that as more specialized laboratory values are introduced into the prediction setting, much higher levels of missingness may be present, and Random Forest continues to perform well for missing data. Our findings on the poor performance of case-wise deletion as an approach to handling missing data are consonant with mainstream recommendations for more than two decades[32].

Our findings on missing data are of note because of the distinctive, yet real-world, way in which missing data were generated. There were two missingness processes. First, clinicians in routine practice only sometimes order any given laboratory, and thus the presence or absence of an order may itself provide prognostic importance. [35] Second, a given effort to identify all of a given target laboratory values may or may not succeed. Even in a large system with a strong tradition of centralization, the extent to which laboratory ascension and labeling practices coincide with their aspiration varies over time, and often clinical insight is necessary to distinguish valid laboratory tests[36]. For any given data pull, it is not trivial to understand which missing values represent failure to find data that exist, versus representing true missingness.

The finding of poorer discrimination of Random Forest in models where the data were fully standardized and cleaned was not anticipated given past literature. The APACHE score was designed to simplify the lab results and to help doctors to predict mortality by hand[2]. Even in its more recent incarnations, APACHE transforms continuous lab results into discrete acute physiology scores[37]. Our data suggests that transforming lab results to APACHE scores is not necessary for Random Forest and may even lead to the loss of information[23]. Remarkably, even standardization to equivalent units across institutions may not be necessary—but at the same time, this means that sources of variance other than simply the laboratory value may also be subtly incorporated into risk-prediction with non-standardized ways. It is a use-case-specific decision as to whether incorporation of such variance is helpful for a given task or is a source of bias.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### Implications

Our findings have implications for both practitioners seeking to implement a given prediction rule and scientists interested in risk-prediction generally. For practitioners, no given method yields consistently superior results in terms of discrimination. Therefore, other performance desiderata, whether psychometric or implementation ease, may play an important role. They also suggest that missing data imputation approaches other than case-wise deletion during development are mandatory.

Our results also note that Random Forests and neural networks were strikingly robust to even guite naively prepared data, in contrast to logistic regression. This suggests that the truth of the oft-quoted aphorisms about "garbage in, garbage out" may depend on the categorization model and missing data imputation method used. In situations where ascertainment and cleaning of data are more costly, random forests may offer pragmatic advantages if these findings .Z.C.Z are replicable.

#### **Strengths and Limitations**

Strengths of our analysis include its use of real world data, with real world data generation and missingness-generation problems on a canonical real world problem. We also used multiple methods implemented in standardized ways. The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility.

Limitations of our analysis stem fundamentally from the nearly infinite combinations of analysis factors that might be varied, and our inability to explore such a high dimensional space. Thus we only considered one outcome, and only considered one standardized and decided upon a priori approach within each dataset / categorization model /

#### **BMJ** Open

missingness imputation triad. Other outcomes may yield different answers. We focus on discrimination, as measured by AUC, but other measurement properties are assuredly also important. And we focused on individual-level prediction, as opposed to considering the impact on hospital-level quality assessment or other tasks for which these results may be

used.

#### CONCLUSION

In sum, our results suggest that while there is little variation in discrimination among alternative statistical classification models in well-cleaned data using modern missing data imputation techniques, there may be important variation across models in real world situations. If these findings are replicated in other data with other outcomes, they may help inform 

pragmatic model selection.

**Figure Captions** 

Figure 1. AUC Scores, Full Model 

Figure 2. AUC Scores for lab-only predictors

Figure 3. Partial Dependence Plots for pH 

Figure 4. Partial Dependence Plots for PaO2

# ACKNOWLEDGEMENTS

#### Author Contributions

**Theodore J. Iwashyna:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing

Cheng Ma: Formal analysis, software, visualization, writing-original draft, and writing-review & editing

Xiao Qing Wang: Data curation; Writing – original draft; Writing – review & editing

Sarah Seelye : Writing – original draft; Writing – review & editing

Ji Zhu : Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

Akbar K. Waljee: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

**Declarations of funding interests:** This work was supported by a career development grant award [CDA 11-217 to A.K.W] and a Merit Review Award Number [IIR 16-024 to A.K.W, 17-045 to T.J.I.] from the United States Department of Veterans Affairs Health Services Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Michigan, the Veterans Affairs, the U.S. Government, or the National Institutes of Health.

**Disclosures:** The authors declare there are no conflicts to disclose

Guarantor of the article: Akbar K. Waljee the submission's guarantor, takes responsibility for the integrity of the work as
 a whole, from inception to published article.

#### Licensing Statement:

I, Akbar Waljee, as the Submitting Author have the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our licence. 

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC),

#### **BMJ** Open

the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

#### **Data Sharing Statement:**

Appendices and statistical code are available via Github at https://github.com/CCMRcodes/GIVO. The dataset cannot be disseminated due to inclusion of sensitive patient information under VA regulations.

<text>

### REFERENCES

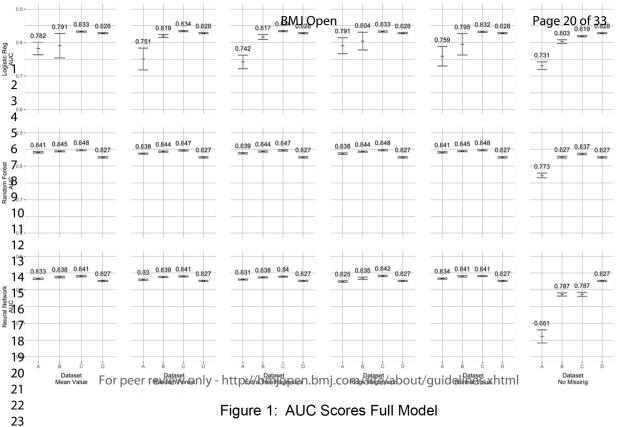
1

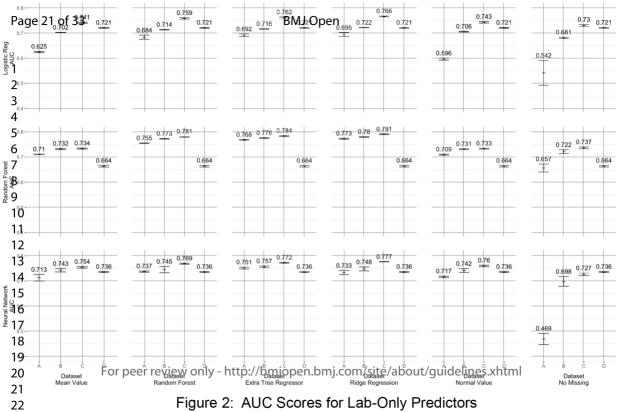
- [1] Iezzoni LI. Risk adjustment for measuring health care outcomes. 4th ed. Chicago, Ill. Arlington, VA: Health
   Administration Press; AUPHA; 2013.
- [2] Lane-Fall MB, Neuman MD. Outcomes measures and risk adjustment. Int Anesthesiol Clin. 2013;51:10-21.
   [2] Outlite Afulas, Part II, Intro duction to Macaures of Outlite (continued), Particular MD.
- [3] Quality AfHRa. Part II. Introduction to Measures of Quality (continued). Rockville, MD.
- [4] Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, et al. Efficacy and safety of
- recombinant human activated protein C for severe sepsis. N Engl J Med. 2001;344:699-709.
- [5] Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction.
   Stat Med. 1984;3:143-52.
- [6] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions
   and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
- [7] van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10
   events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163.
- 16 [8] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a
- multivariable prediction model: PART II binary and time-to-event outcomes. Stat Med. 2019;38:1276-96.
- [9] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic
   prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28:2455-74.
- 20 [10] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a
- multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern
   Med. 2015;162:W1-73.
- [11] Steyerberg EW, ProQuest (Firm). Clinical prediction models a practical approach to development, validation, and
   updating. New York: Springer; 2009. p. xxviii, 497 p.
- 25 Industrig: New York. Springer, 2009. p. XXVII, 497 p.
   26 [12] Render ML, Kim HM, Welsh DE, Timmons S, Johnston J, Hui S, et al. Automated intensive care unit risk adjustment:
   27 results from a National Veterans Affairs study. Crit Care Med. 2003;31:1638-46.
- [13] Render ML, Deddens J, Freyberg R, Almenoff P, Connors AF, Wagner D, et al. Veterans Affairs intensive care unit risk adjustment model: validation, updating, recalibration. Crit Care Med. 2008;36:1031-42.
- 30 [14] Render ML, Freyberg RW, Hasselbeck R, Hofer TP, Sales AE, Deddens J, et al. Infrastructure for quality
- transformation: measurement and reporting in veterans administration intensive care units. BMJ Qual Saf. 2011;20:498 507.
- [15] Wang XQ, Vincent BM, Wiitala WL, Luginbill KA, Viglianti EM, Prescott HC, et al. Veterans Affairs patient database
   (VADD 2014 2017) he illing and in a statement of the statement o
- (VAPD 2014-2017): building nationwide granular data for clinical discovery. BMC Med Res Methodol. 2019;19:94.
- [16] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying
   laboratory observations: a 5-year update. Clin Chem. 2003;49:624-33.
- [17] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and
   codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results.
   Clin Chem 1996:42:81-90
- Clin Chem. 1996;42:81-90.
   [18] van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures
- <sup>42</sup> into a point system for hospital death using administrative data. Med Care. 2009;47:626-33.
- 43
  43
  43
  [19] HCUP-US Tools & Software Page. 2019.
- [20] Hooper TI, Gackstetter GD, Leardmann CA, Boyko EJ, Pearse LA, Smith B, et al. Early mortality experience in a large
   military cohort and a comparison of mortality data sources. Popul Health Metr. 2010;8:15.
- 47 [21] Prescott HC, Kepreos KM, Wiitala WL, Iwashyna TJ. Temporal Changes in the Influence of Hospitals and Regional
- 48 Healthcare Networks on Severe Sepsis Mortality. Crit Care Med. 2015;43:1368-74.
- <sup>49</sup> [22] Breiman L. Classification and regression trees. New York, NY: Chapman & Hall; 1993.
- <sup>50</sup> [23] Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
- [24] Omidvar O, Dayhoff JE, ScienceDirect (Online service). Neural networks and pattern recognition. San Diego, Calif.:
   Academic Press; 1998.
- [25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python.
   Journal of machine learning research. 2011;12:2825–30.
- 56
- 57
- 58 59

60

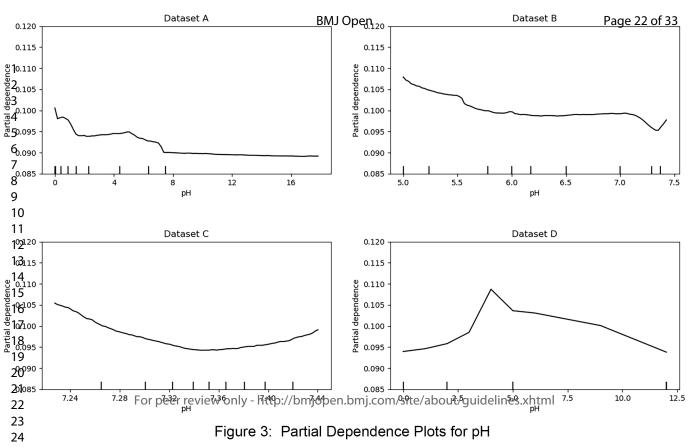
#### **BMJ** Open

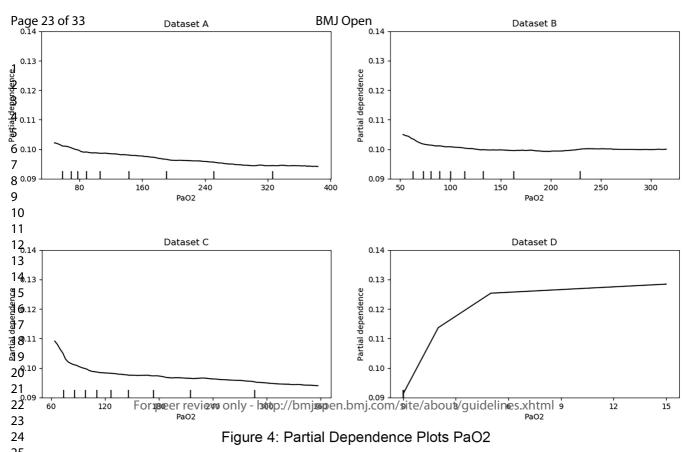
- [26] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3.
- [27] Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data.
- Bioinformatics. 2012;28:112-8.
- [28] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63:3-42.
- [29] Maree R, Geurts P, Wehenkel L. Random subwindows and extremely randomized trees for image classification in
- cell biology. BMC Cell Biol. 2007;8 Suppl 1:S2.
- [30] Hastie T, Friedman J, Tibshirani R, SpringerLink (Online service). The Elements of Statistical Learning Data Mining, Inference, and Prediction. New York, NY: Springer New York : Imprint: Springer; 2001.
- [31] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health
- evaluation: a physiologically based classification system. Crit Care Med. 1981;9:591-7.
- [32] Allison PD, Sage Publications. Missing data. Thousand Oaks, [Calif.] ; London: SAGE; 2002. p. 1 online resource (vi, 91 p.).
- [33] Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees.
- Proceedings of the 26th International Conference on Neural Information Processing Systems Volume 1. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. p. 431-9.
- [34] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001;29:1189-232.
- [35] Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.
- [36] Wiitala WL, Vincent BM, Burns JA, Prescott HC, Waljee AK, Cohen GR, et al. Variation in Laboratory Test Naming
- Conventions in EHRs Within and Between Hospitals: A Nationwide Longitudinal Study. Med Care. 2019;57:e22-e7.
- [37] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV:
- hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34:1297-310. terez onz





~ ~





#### Appendix A. Patient-level variables included in models

Demographics	Gender, Age, Race (White, Black or African American, Asian, Native Hawaiian or other Pacific Islander, Unknown), Hispanic ethnicity
Comorbidities, included in Elixhauser	Hypertension, Congestive Heart Failure, Cardiac Arrhythmia, Valvular Disease, Pulmonary Circulation Disorders, Peripheral Vascular Disorders, Paralysis, Other Neurological Disorders, Chronic Pulmonary Disease, Diabetes Uncomplicated, Diabetes Complicated, Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer Disease excluding bleeding, AIDS/HIV, Lymphoma, Metastatic Cancer, Solid Tumor without Metastasis, Rheumatoid Arthritis/Collagen, Coagulopathy, Obesity, Weight Loss, Fluid and Electrolyte Disorders, Blood Loss Anemia, Deficiency Anemia, Alcohol Abuse, Drug Abuse, Psychoses, Depression
Diagnoses, HCUP CCS single-level and multi-level	Top 20 most frequent single-level CCS diagnoses: Congestive Heart Failure (non-hypertensive), Non-specific Chest Pain, Coronary Atherosclerosis and Other Heart Disease, Cardiac Dysrhythmias, Alcohol-related Disorders, Septicemia (except in labor), Chronic Obstructive Pulmonary Disease and Bronchiectasis, Pneumonia, Skin and Subcutaneous Tissue Infections, Osteoarthritis, Complication of Device (implant or graft), Complications of Surgical Procedures or Medical Care, Diabetes Mellitus with Complications, Respiratory Failure, Urinary Tract Infections, Renal Failure, Spondylosis, Acute Myocardial Infarction, Fluid and Electrolyte Disorders, Gastrointestinal Hemorrhage 18 level 1 multi-level CCS categories: Infectious and Parasitic Diseases, Neoplasms, Endocrine Disorders, Anemia, Mental Illness, Diseases of the Nervous System, Diseases of the Circulatory System, Diseases of the Respiratory System, Diseases of the Digestive System, Diseases of the Genitourinary System, Complications of Pregnancy or Childbirth, Skin Disease, Diseases of the Musculoskeletal System, Congenital Anomalies, Perinatal Conditions, Injury and Poisoning, Other Health Status Conditions, Other Residual Codes
Laboratory values	Albumin, Bilirubin, Blood Urea Nitrogen, Creatinine, Glucose, Hematocrit, Partial pressure of oxygen score, pH score, Sodium, White Blood Cell

#### Appendix B

Table B.1: Model Performances (Full Model)

Classification		Imputation	AUROC	Optimal	Predict	ed Cases	Accurat	te Rate			Brier
Method	Dataset	Method	(95%CI)	Cutoff	Death	Survival	Death	Survival	Sensitivity	Specificity	Score
Logistic Regression	A	Mean Value	0.78(0.76- 0.80)	0.48	37199	69549	0.21	0.96	0.74	0.69	0.19
Logistic Regression	В	Mean Value	0.79(0.75- 0.83)	0.46	34915	71833	0.24	0.97	0.80	0.72	0.17
Logistic Regression	С	Mean Value	0.83(0.83- 0.83)	0.46	35970	70778	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Random Forest	0.75(0.72- 0.78)	0.49	39528	67220	0.18	0.95	0.69	0.66	0.21
Logistic Regression	В	Random Forest	0.82(0.82- 0.82)	0.49	31996	74752	0.25	0.97	0.77	0.75	0.17
Logistic Regression	С	Random Forest	0.83(0.83- 0.84)	0.46	36017	70731	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Extra Trees Regression	0.74(0.72- 0.76)	0.46	45762	60986	0.17	0.96	0.74	0.61	0.21
Logistic Regression	В	Extra Trees Regression	0.82(0.81- 0.82)	0.47	33642	73106	0.25	0.97	0.79	0.74	0.17
Logistic Regression	С	Extra Trees Regression	0.83(0.83- 0.84)	0.47	34579	72169	0.24	0.97	0.78	0.73	0.17
Logistic Regression	A	Ridge Regression	0.79(0.77- 0.82)	0.46	38579	68169	0.21	0.97	0.78	0.68	0.18
Logistic Regression	В	Ridge Regression	0.80(0.78- 0.83)	0.46	35034	71714	0.24	0.97	0.80	0.72	0.17

Page	26	of	33

Logistic		Ridge	0.83(0.83-	0.47	35220	71528	0.23	0.97	0.78	0.72	0.1
Regression	С	Regression	0.84)	••••	00110	/ _0_0	0.20	0107		0.7 -	0.1
Logistic		Normal	0.76(0.73-	0.50	37392	69356	0.18	0.95	0.63	0.68	0.2
Regression	A	Value	0.79)	0.50	57552	05550	0.10	0.55	0.05	0.00	0.2
Logistic		Normal	0.80(0.76-	0.50	30977	75771	0.26	0.97	0.76	0.76	0.1
Regression	В	Value	0.83)	0.50	30977	/5//1	0.20	0.97	0.70	0.70	0.1
Logistic		Normal	0.83(0.83-	0.46	35676	71072	0.23	0.97	0.78	0.72	0.1
Regression	С	Value	0.83)	0.40	33070	/10/2	0.25	0.57	0.78	0.72	0.1
Logistic			0.73(0.72-	0.43	37658	69090	0.18	0.95	0.66	0.68	0.1
Regression	A	No Missing	0.74)	Ű	0,000	23030	0.10	0.00	0.00	0.00	0.1
Logistic			0.8(0.80-	0.42	33153	73595	0.24	0.97	0.76	0.74	0.1
Regression	В	No Missing	0.81)	0.12	00100		0.2 .	0.07	0.70	0.7.1	0.1
Logistic			0.82(0.82-	0.44	35333	71415	0.22	0.96	0.76	0.72	0.1
Regression	С	No Missing	0.82)	••••		C		0100		0.7 -	0.1
Logistic			0.83(0.83-	0.47	34184	72564	0.24	0.97	0.78	0.73	0.1
Regression	D	None	0.83)	0.17	01101	, 2001		0.57	0.70	0170	0.1
Random		Mean	0.84(0.84-	0.12	32330	74418	0.26	0.97	0.79	0.75	0.0
Forest	A	Value	0.84)	0.12	52555	,	0.20	0.57		0175	0.0
Random		Mean	0.85(0.84-	0.11	32642	74106	0.26	0.97	0.80	0.75	0.0
Forest	В	Value	0.85)	0.11	52012	, 1200	0.20	0.57	0.00		0.0
Random		Mean	0.85(0.85-	0.11	33548	73200	0.25	0.97	0.81	0.74	0.0
Forest	С	Value	0.85)	0.11	33340	, 5200	0.25	0.07	0.01	0.7 4	0.0
Random		Random	0.84(0.84-	0.12	32659	74089	0.25	0.97	0.78	0.75	0.0
Forest	А	Forest	0.84)	0.12	52055	, 4005	0.25	0.57	0.76	0.75	0.0
Random		Random	0.84(0.84-	0.11	34093	72655	0.25	0.97	0.81	0.73	0.0
Forest	В	Forest	0.85)	0.11	34033	12000	0.25	0.97	0.01	0.75	0.0

Page 27 of 33

Random		Random	0.85(0.85-								
Forest	С	Forest	0.85)	0.11	33029	73719	0.25	0.97	0.80	0.74	0.07
Random		Extra Trees	0.84(0.84-	0.11	32938	73810	0.25	0.97	0.79	0.74	0.07
Forest	А	Regression	0.84)	0.11	52550	/3010	0.25	0.57	0.75	0.74	0.07
Random		Extra Trees	0.84(0.84-	0.12	32411	74337	0.26	0.97	0.80	0.75	0.07
Forest	В	Regression	0.85)	0.12	52411	74337	0.20	0.57	0.00	0.75	0.07
Random		Extra Trees	0.85(0.85-	0.11	33567	73181	0.25	0.97	0.80	0.74	0.07
Forest	С	Regression	0.85)	0.11	55507	/5101	0.25	0.57	0.80	0.74	0.07
Random		Ridge	0.84(0.45-	0.11	34587	72161	0.24	0.97	0.80	0.73	0.07
Forest	А	Regression	0.84)		54507	/2101	0.24	0.57	0.00	0.75	0.07
Random		Ridge	0.84(0.84-	0.12	31643	75105	0.26	0.97	0.79	0.76	0.07
Forest	В	Regression	0.85)	0.12	510.15	, 5105	0.20	0.57	0.75	0.70	0.07
Random		Ridge	0.85(0.85-	0.12	32531	74217	0.25	0.97	0.79	0.75	0.07
Forest	С	Regression	0.85)	0.12	02001		0.25	0.57	0.75	0.75	0.07
Random		Normal	0.84(0.84-	0.12	31234	75514	0.26	0.97	0.78	0.76	0.07
Forest	A	Value	0.84)	0.12	01201	,	0.20	0.57	0.70	0.70	0.07
Random		Normal	0.85(0.84-	0.11	32711	74037	0.26	0.97	0.80	0.75	0.07
Forest	В	Value	0.85)	0.11	52711	74037	0.20	0.57		0.75	0.07
Random		Normal	0.85(0.85-	0.12	31159	75589	0.26	0.97	0.78	0.76	0.07
Forest	С	Value	0.85)	0.12	51155	/ 3303	0.20	0.57	0.70	0.70	0.07
Random			0.77(0.77-	0.18	36332	70416	0.20	0.96	0.71	0.70	0.08
Forest	A	No Missing	0.78)	0.10	30332	,0410	0.20	0.50	0.7 1	0.70	0.00
Random			0.83(0.82-	0.16	31836	74912	0.25	0.97	0.77	0.75	0.07
Forest	В	No Missing	0.83)	0.10	51050	, 7, 7, 7, 2, 2, 2	0.25	0.07	0.77	0.75	0.07
Random			0.84(0.83-	0.16	34517	72231	0.24	0.97	0.78	0.73	0.08
Forest	С	No Missing	0.84)	0.10	7471/	12231	0.24	0.97	0.78	0.75	0.08

**BMJ** Open

Page	28	of	33

Random			0.83(0.83-	0.11	33407	73341	0.24	0.97	0.77	0.74	0.0
Forest	D	None	0.83)	0.11	33407	75541	0.24	0.57	0.77	0.74	0.0
Neural		Mean	0.83(0.83-	0.53	35031	71717	0.24	0.97	0.80	0.72	0.1
Network	А	Value	0.84)	0.55	33031	/1/1/	0.24	0.97	0.80	0.72	0.1
Neural		Mean	0.84(0.84-	0.52	22716	74022	0.25	0.97	0.79	0.75	0.1
Network	В	Value	0.84)	0.52	32716	74032	0.25	0.97	0.79	0.75	0.1
Neural		Mean	0.84(0.84-	0.52	225.40	74100	0.25	0.07	0.70	0.75	0.1
Network	С	Value	0.84)	0.53	32549	74199	0.25	0.97	0.79	0.75	0.1
Neural		Random	0.83(0.83-	0.55	32515	74233	0.25	0.97	0.77	0.75	0.1
Network	А	Forest	0.83)	0.55	52515	74233	0.25	0.57	0.77	0.75	0.1
Neural		Random	0.84(0.84-	0.57	30842	75906	0.26	0.97	0.77	0.76	0.1
Network	В	Forest	0.84)	0.57	50042	75500	0.20	0.57	0.77	0.70	0
Neural		Random	0.84(0.84-	0.55	34144	72604	0.24	0.97	0.79	0.73	0.1
Network	С	Forest	0.84)	0.55	34144	72004	0.24	0.57	0.75	0.75	0.1
Neural		Extra Trees	0.83(0.83-	0.50	37351	69397	0.23	0.97	0.82	0.70	0.1
Network	А	Regression	0.83)	0.50	57551	05557	0.25	0.57	0.02	0.70	0.1
Neural		Extra Trees	0.84(0.84-	0.54	33529	73219	0.25	0.97	0.80	0.74	0.1
Network	В	Regression	0.84)	0.54	33323	75215	0.25	0.57	0.80	0.74	0.1
Neural		Extra Trees	0.84(0.84-	0.49	33324	73424	0.25	0.97	0.79	0.74	0.1
Network	С	Regression	0.84)	0.45	33324	73424	0.25	0.57	0.75	0.74	0.1
Neural		Ridge	0.83(0.82-	0.51	33864	72884	0.24	0.97	0.78	0.73	0.1
Network	А	Regression	0.83)	0.51	55004	/2004	0.24	0.97	0.76	0.75	0.1
Neural		Ridge	0.84(0.83-	0.52	31186	75562	0.26	0.97	0.78	0.76	0.1
Network	В	Regression	0.84)	0.52	21100	10002	0.20	0.97	0.76	0.70	0.1
Neural		Ridge	0.84(0.84-	0.55	32145	74603	0.25	0.97	0.78	0.75	0.1
Network	С	Regression	0.84)	0.55	32143	74003	0.25	0.97	0.76	0.75	0.1

Page	29	of	33
------	----	----	----

 BMJ Open

Neural Network	А	Normal Value	0.83(0.83- 0.84)	0.58	31675	75073	0.25	0.97	0.77	0.76	0.19
Neural Network	В	Normal Value	0.84(0.84- 0.84)	0.45	35026	71722	0.24	0.97	0.81	0.72	0.16
Neural Network	С	Normal Value	0.84(0.84- 0.84)	0.55	32864	73884	0.25	0.97	0.79	0.75	0.18
Neural Network	А	No Missing	0.66(0.64- 0.68)	0.76	49011	57737	0.14	0.94	0.65	0.56	0.50
Neural Network	В	No Missing	0.79(0.78- 0.79)	0.59	32676	74072	0.23	0.96	0.71	0.74	0.21
Neural Network	С	No Missing	0.79(0.78- 0.79)	0.59	38619	68129	0.20	0.96	0.75	0.68	0.24
Neural Network	D	None	0.83(0.83- 0.83)	0.52	33760	72988	0.24	0.97	0.78	0.73	0.18
						-	6	4	0	4	

#### Classification Imputation Predicted Cases AUROC Optimal Accurate Rate Brier Method (95%CI) Survival Dataset Method Cutoff Death Survival Death Sensitivity Specificity Score Logistic 0.63(0.62-Mean 0.50 32327 74421 0.16 0.93 0.50 0.72 0.24 0.63) Regression А Value 0.70(0.70-Logistic Mean 0.47 73398 0.21 0.95 0.73 33350 0.66 0.20 Regression В Value 0.70) Logistic 0.74(0.74-Mean 0.47 35248 71500 0.19 0.95 0.62 0.70 0.22 0.74) Regression С Value 0.68(0.68-Logistic Random 0.48 39566 67182 0.17 0.94 0.63 0.66 0.23 0.69) Regression А Forest 0.71(0.71-Logistic Random 37758 68990 0.45 0.20 0.96 0.71 0.69 0.20 Regression В Forest 0.71) 0.76(0.76-Logistic Random 0.47 68327 0.18 38421 0.95 0.66 0.67 0.21 Regression С Forest 0.76) Logistic Extra Trees 0.69(0.69-0.46 43607 63141 0.17 0.95 0.69 0.62 0.22 Regression А Regression 0.70) Logistic 0.72(0.72-Extra Trees 0.20 0.73 0.44 39295 67453 0.96 0.67 0.19 0.72) Regression В Regression 0.76(0.76-Logistic Extra Trees 0.45 64073 0.71 0.63 42675 0.17 0.95 0.21 Regression С Regression 0.76) 0.70(0.69-Logistic Ridge 0.47 64234 0.17 0.95 0.69 0.63 0.22 42514 0.70) Regression А Regression Logistic Ridge 0.72(0.72-0.44 39856 66892 0.20 0.96 0.75 0.67 0.19 0.72) Regression В Regression

#### Table B.2: Model Performance (Using only lab variables)

1 2 3

> 4 5

> 6

7 8

9

10 11

12

13 14

15

16 17

18

19 20

21

22 23

24

25 26

27

28 29

30

31

32

33

34

35

36 37

38

45 46 47

Page 31 o	of 33
-----------	-------

Logistic		Ridge	0.77(0.76-	0.45	40707	C 4 0 4 4	0.47	0.05	0.74	0.62	0.0
Regression	С	Regression	0.77)	0.45	42737	64011	0.17	0.95	0.71	0.63	0.2
Logistic		Normal	0.60(0.59-	0.49	31990	74758	0.15	0.92	0.46	0.72	0.2
Regression	А	Value	0.60)	0.45	51550	/4/30	0.15	0.52	0.40	0.72	0.2
Logistic		Normal	0.71(0.70-	0.45	38325	68423	0.19	0.95	0.70	0.68	0.2
Regression	В	Value	0.71)	0.45	30323	00425	0.15	0.55	0.70	0.00	0.2
Logistic		Normal	0.74 (0.74-	0.46	41447	65301	0.17	0.95	0.68	0.64	0.2
Regression	С	Value	0.75)	0.40	41447	05501	0.17	0.55	0.00	0.04	0.2
Logistic			0.54 (0.49-	0.57	17678	89070	0.15	0.91	0.25	0.84	0.2
Regression	А	No Missing	0.59)	0.57	17078	85070	0.15	0.51	0.25	0.04	0.2
Logistic			0.68 n(0.68-	0.45	32766	73982	0.20	0.95	0.64	0.73	0.1
Regression	В	No Missing	0.68)	0.45	52700	75502	0.20	0.55	0.04	0.75	0.1
Logistic			0.73(0.73-	0.50	30965 <	75783	0.19	0.94	0.55	0.74	0.2
Regression	С	No Missing	0.73)	0.50	30505	13783	0.15	0.54	0.55	0.74	0.2
Logistic			0.72(0.72-	0.49	35766	70982	0.19	0.95	0.64	0.70	0.2
Regression	D	None	0.72)	0.45	33700	70502	0.15	0.55	0.04	0.70	0.2
Random		Mean	0.71(0.71-	0.09	46226	60522	0.17	0.95	0.73	0.60	0.0
Forest	А	Value	0.71)	0.05	40220	00522	0.17	0.55	0.75	0.00	0.0
Random		Mean	0.73(0.73-	0.10	44628	62120	0.18	0.96	0.75	0.62	0.0
Forest	В	Value	0.73)	0.10	44020	02120	0.10	0.50	0.75	0.02	0.0
Random		Mean	0.73(0.73-	0.10	44525	62223	0.18	0.96	0.75	0.62	0.0
Forest	С	Value	0.74)	0.10	77525	02225	0.10	0.50	0.75	0.02	0.0
Random		Random	0.76(0.75-	0.09	41715	65033	0.18	0.96	0.74	0.65	0.0
Forest	А	Forest	0.76)	0.09	-11/1J	00000	0.10	0.90	0.74	0.05	0.0
Random		Random	0.77(0.77-	0.10	38154	68594	0.20	0.96	0.75	0.69	0.0
Forest	В	Forest	0.77)	0.10	30134	00334	0.20	0.90	0.75	0.09	0.0

Random		Random	0.78(0.78-	0.09	40709	66039	0.19	0.96	0.75	0.66	0.08
Forest	С	Forest	0.78)	0.05	40705	00035	0.15	0.50	0.75	0.00	0.00
Random		Extra Trees	0.77(0.77-	0.09	43230	63518	0.19	0.96	0.77	0.64	0.08
Forest	А	Regression	0.77)	0.09	43230	05518	0.19	0.90	0.77	0.04	0.00
Random		Extra Trees	0.78(0.77-	0.10	38734	68014	0.20	0.96	0.76	0.68	0.08
Forest	В	Regression	0.78)	0.10	56754	08014	0.20	0.90	0.70	0.08	0.00
Random		Extra Trees	0.78(0.78-	0.10	38810	67938	0.20	0.96	0.74	0.68	0.08
Forest	С	Regression	0.79)	0.10	20010	07950	0.20	0.90	0.74	0.08	0.00
Random		Ridge	0.77(0.77-	0.10	39913	66835	0.20	0.96	0.75	0.67	0.08
Forest	А	Regression	0.78)	0.10	33313	00000	0.20	0.50	0.75	0.07	0.00
Random		Ridge	0.78(0.78-	0.09	39663	67085	0.20	0.97	0.77	0.67	0.0
Forest	В	Regression	0.78)	0.05	35005	0,003	0.20	0.57	0.77	0.07	0.00
Random		Ridge	0.79(0.79-	0.09	40249	66499	0.20	0.96	0.76	0.66	0.0
Forest	С	Regression	0.79)	0.05	40245	00455	0.20	0.50	0.70	0.00	0.00
Random		Normal	0.71(0.71-	0.10	46047	60701	0.17	0.95	0.73	0.60	0.0
Forest	Α	Value	0.71)	0.10	40047	00701	0.17	0.55	0.75	0.00	0.0.
Random		Normal	0.73(0.73-	0.10	44400	62348	0.18	0.96	0.75	0.62	0.09
Forest	В	Value	0.73)	0.10	44400	02340	0.10	0.50	0.75	0.02	0.0.
Random		Normal	0.73(0.73-	0.09	46774	59974	0.17	0.96	0.77	0.60	0.0
Forest	С	Value	0.74)	0.05	40774	55574	0.17	0.50	0.77	0.00	0.0
Random			0.66(0.64-	0.26	20159	86589	0.21	0.93	0.40	0.83	0.1
Forest	А	No Missing	0.67)	0.20	20133	00009	0.21	0.55	0.40	0.85	0.10
Random			0.72(0.71-	0.14	52201	54547	0.16	0.96	0.78	0.54	0.0
Forest	В	No Missing	0.73)	0.14	52201	54547	0.10	0.90	0.76	0.54	0.00
Random			0.74(0.73-	0.21	26193	80555	0.22	0.94	0.55	0.79	0.0
Forest	С	No Missing	0.74)	0.21	20192	00000	0.22	0.94	0.55	0.75	0.03

Page	33	of	33
------	----	----	----

Random			0.66(0.66-	0.45	33868	72880	0.18	0.94	0.57	0.71	0
Forest	D	None	0.67)	0.45	55000	72000	0.10	0.54	0.57	0.71	U
Neural		Mean	0.71(0.70-	0.47	33169	73579	0.19	0.94	0.60	0.72	0
Network	A	Value	0.73)	0.47	33103	13313	0.15	0.54	0.00	0.72	Ū
Neural		Mean	0.74(0.74-	0.49	31171	75577	0.21	0.95	0.63	0.75	0
Network	В	Value	0.75)	0.45	511/1	13311	0.21	0.55	0.05	0.75	0
Neural		Mean	0.75(0.75-	0.40	42096	64652	0.18	0.96	0.74	0.64	0
Network	С	Value	0.76)	0.40	42090	04032	0.10	0.90	0.74	0.04	0
Neural		Random	0.74(0.74-	0.50	37930	68818	0.19	0.95	0.68	0.68	0
Network	А	Forest	0.74)	0.50	57550	00010	0.15	0.55	0.00	0.00	U
Neural		Random	0.75(0.73-	0.54	36554	70194	0.20	0.96	0.71	0.70	0
Network	В	Forest	0.76)	0.54	50554	70154	0.20	0.50	0.71	0.70	0
Neural		Random	0.77(0.77-	0.42	42444	64304	0.18	0.96	0.74	0.64	0
Network	C	Forest	0.77)	0.42	72777	04304	0.10	0.50	0.74	0.04	0
Neural		Extra Trees	0.75(0.75-	0.42	44585	62163	0.18	0.96	0.76	0.62	0
Network	А	Regression	0.76)	0.42	4505	02105	0.10	0.50	0.70	0.02	0
Neural		Extra Trees	0.76(0.75-	0.46	40067	66681	0.20	0.96	0.75	0.67	0
Network	В	Regression	0.76)	0.40	40007	00081	0.20	0.90	0.75	0.07	0
Neural		Extra Trees	0.77(0.77-	0.48	39088	67660	0.19	0.96	0.71	0.67	0.
Network	С	Regression	0.77)	0.40	39088	07000	0.19	0.90	0.71	0.07	0
Neural		Ridge	0.73(0.73-	0.49	43129	63619	0.18	0.96	0.74	0.63	0
Network	А	Regression	0.74)	0.49	43129	02013	0.10	0.90	0.74	0.05	0
Neural		Ridge	0.75(0.74-	0.44	39388	67360	0.20	0.96	0.74	0.67	0.
Network	В	Regression	0.76)	0.44	39388	0/300	0.20	0.90	0.74	0.67	0
Neural		Ridge	0.78(0.78-	0.52	20040	60700	0.10	0.05	0.69	0.69	
Network	С	Regression	0.78)	0.53	38048	68700	0.19	0.95	0.68	0.68	0

Neural		Normal	0.72(0.71-	0.45	33193	73555	0.19	0.95	0.61	0.72	0.18
Network	A	Value	0.72)								
Neural		Normal	0.74(0.73-	0.50	37560	69188	0.20	0.96	0.71	0.69	0.20
Network	В	Value	0.75)	0.00	0,000	09100	0.20	0.50	0.71	0.05	0.20
Neural		Normal	0.76(0.76-	0.41	10107	64561	0.18	0.96	0.74	0.64	0.18
Network	С	Value	0.76)	0.41	42107	04501	0.18	0.90	0.74	0.04	0.10
Neural			0.47(0.45-	0.79	2628	104120	0.14	0.90	0.04	0.98	0.44
Network	А	No Missing	0.49)	0.79	2020	104120	0.14	0.90	0.04	0.98	0.44
Neural			0.70(0.68-	0.69	22720	83028	0.23	0.94	0.52	0.81	0.31
Network	В	No Missing	0.72)	0.09	23720	83028	0.25	0.94	0.52	0.81	0.51
Neural			0.73(0.72-	0.57	51077	55671	0.15	0.95	0.74	0.55	0.29
Network	С	No Missing	0.73)	0.57	510/7	55071	0.15	0.55	0.74	0.55	0.25
Neural			0.74(0.73-	0.50	36925	69823	0.19	0.95	0.67	0.69	0.21
Network	D	None	0.74)	0.50	30323	03023	0.15	0.55	0.07	0.05	0.21
						L.	10				

# **BMJ Open**

#### Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU Mortality, a US Nationwide Retrospective Cohort and Simulation Study

Journal:	BMJ Open
Manuscript ID	bmjopen-2020-041421.R1
Article Type:	Original research
Date Submitted by the Author:	25-Aug-2020
Complete List of Authors:	Iwashyna, Theodore; University of Michigan, Internal Medicine; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Ma, Cheng; University of Michigan, Statistics Wang, Xiao Qing; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Seelye, Sarah; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Zhu, Ji; University of Michigan, Department of Statistics Waljee, Akbar; University of Michigan, Gastroenterology; VA Center for Clinical Management Research, Ann Arbor VA Medical Center
<b>Primary Subject Heading</b> :	Research methods
Secondary Subject Heading:	Health services research, Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, STATISTICS & RESEARCH METHODS, Adult intensive & critical care < INTENSIVE & CRITICAL CARE

SCHOLARONE<sup>™</sup> Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

reliez oni

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Title: Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU ماريمينا والمربية المربية

1 2 3	Mortality, a US Nationwide Retrospective Cohort and Simulation Study
4 5	Authors:
6 7	Theodore J. Iwashyna*, M.D., PhD (1,2,3)
8 9	Cheng Ma*, B.S. (4)
10 11	Xiao Qing Wang, MPH (1)
12	Sarah Seelye, Ph.D. (1)
13 14	Ji Zhu, Ph.D. (4)
15 16	Akbar K. Waljee, M.D., M.Sc. (1,2,3)
17 18 19 20 21	<u>*considered co-first authors.</u>
22 23 24 25 26 27	<ol> <li>VA Center for Clinical Management Research, VA Ann Arbor Health Care System, Ann Arbor, MI, USA.</li> <li>Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA.</li> <li>Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), Ann Arbor, MI, USA</li> <li>Department of Statistics, University of Michigan, Ann Arbor, MI, USA.</li> </ol>
28 29 30 31 32	Address correspondence to: Akbar K. Waljee, M.D. M.Sc., Associate Professor of Medicine, Division of Gastroenterology, Department of Internal Medicine, Michigan Medicine, Ann Arbor Veterans Affairs Medical Center, 2215 Fuller Road, 111D, Ann Arbor, Michigan 48105
33 34	Phone: 734-845-5865; Fax: 734-845-3091; E-mail: <u>awaljee@med.umich.edu</u>
35	
36 37 38	Keywords: missing data, risk prediction, machine learning, electronic health record data, random forests
39 40	Word count: 3,259
41 42	
43	
44 45	
46 47	
48	
49 50	
51	
52 53	
53 54	
55 56	
57	
58 59	1
60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### BMJ Open

#### ABSTRACT

2 3	<b>Objective:</b> There has been a proliferation of approaches to statistical methods and missing data imputation as electronic
4 5 6	health records become more plentiful; however, the relative performance on real-world problems is unclear.
7 8	Materials and Methods: Using 355,823 ICU hospitalizations at over 100 hospitals in the nationwide VA healthcare
9 10 11	system (2014-2017), we systematically varied 3 approaches: how we extracted and cleaned physiologic variables; how
12 13	we handled missing data (using mean value imputation, random forest, extremely randomized tress (extra-trees
14 15	regression), ridge regression, normal value imputation, and case-wise deletion); and how we computed risk (using
16 17 18	logistic regression, random forest, and neural networks). We applied these approaches in a 70% development sample
19 20	and tested the results in an independent 30% testing sample. Area under the ROC Curve (AUROC) was used to quantify
21 22	model discrimination.
23 24 25	Results: In 355,823 ICU stays, there were 34,867 deaths (9.8%) within 30 days of admission. The highest AUROC's
26 27	obtained for each primary classification method were very similar: 0.83 (95% CI [0.83-0.83]) to 0.85 (95% CI 0.840.85).
28 29	Likewise, there was relatively little variation within classification method by the missing value imputation method
30 31 32	used—except when case-wise deletion was applied for missing data.
33 34	<b>Conclusion:</b> Variation in discrimination was seen as a function of data cleanliness, with logistic regression suffering the
35 36 37	most loss of discrimination in the least clean data. Losses in discrimination were not present in random forest and neural
38 39	networks even in naively extracted data. Data from a large nationwide health system revealed interactions between
40 41 42	missing data imputation techniques, data cleanliness, and classification methods for predicting 30-day mortality.
42 43 44	
45 46	
47 48 49	
50 51	
52 53	
54 55 56	
57 58	2
59 60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study focuses on a large, real world dataset consisting of 355,823 ICU stays at over 100 different facilities. •
- Multiple methods of model fitting and missing data imputation were implemented in standardized ways that . reflect common practice.
- The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility, and we encourage validation on other datasets.
- Due to high dimensionality of method combinations, this study only considered one outcome, and only . considered one standardized and decided upon an a priori approach within each dataset / categorization model / missingness imputation triad.

Intainon is uvalidation os od combinations, is decided upon an a ps .

#### **BMJ** Open

#### INTRODUCTION

Risk adjustment plays an increasingly central role in the organization, care of, and science about critically ill patients[1, 2]. Statistical adjustment, including the handling of missing data, is essential for many performance measurements as well as pay-for-performance and shared savings systems. It is used to stratify the care of patients for treatments and track quality improvement efforts over time[3]. It is routinely measured, even in clinical trials, to assess confounder balance between arms and may form part of RCT enrollment or drug approval criteria[4].

As a result, there has been a proliferation of risk scores and missing data imputation tools both for the common task of short-term mortality prediction and for more specialized tasks. Many statistical tools have been promoted. Rules of thumb have developed and existed long enough to be critiqued[5-9]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines offer standardization of reporting[10]. Textbooks have emerged[11]. Yet questions remain on fundamental pragmatic issues: How clean does the data have to be to prevent the so-called "garbage in, garbage out (GIGO)" phenomenon? How sensitive are methods to missing data and how should it be handled? Do these analytic decisions interact?

To address such questions, we compared the performance of an array of methods on a single standardized problem—the prediction of 30-day mortality based ondemographics, day 1 laboratory results, comorbidities, and diagnoses among patients admitted to the Intensive Care Unit (ICU) at any hospital in the nationwide Veterans Health Administration system[12-14]. Using the same set of real ICU admissions, we systematically varied three parameters: the approach used to extract and clean physiologic variables from the electronic health record; the approach used to handle missing data; and the approach used to compute the risk. We systematically applied these approaches in a 70% development sample and tested the results in an independent 30% testing sample, to provide real world comparisons to inform future pragmatic implementation of risk scores.

#### METHODS

#### Cohort

Data were drawn from the Veterans Affairs Patient Database (VAPD), which contains daily patient physiology for acute hospitalizations between January 1, 2014 and December 31, 2017. The VAPD includes patient demographics, laboratory results, and diagnoses that are commonly used to predict 30-day mortality from the day of admission. Here, we included data from all ICU hospitalizations on day 1 of each hospitalization. Full details of the VAPD have been published elsewhere[15].

The development of this database was reviewed and approved by the VA Ann Arbor Healthcare System's Institutional Review Board.

Four versions of the dataset were created for each hospitalization on admission: A) raw lab values extracted using only lab test names, B) raw lab values extracted using only Logical Observation Identifiers Names and Codes (LOINC), C) cleaned lab values extracted using both LOINC[16, 17] and searched text lab test names, and D) cleaned lab values converted to Acute Physiology And Chronic Health Evaluation (APACHE) points, extracted using both LOINC and lab test names.

#### No Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

#### **Predictor Variables**

In our primary analyses, we adjust for 10 laboratory values that were collected within one day of hospital admission. Further patient-level adjustments included demographic characteristics (gender, age, race, and Hispanic ethnicity), 30 comorbidities, and 38 primary diagnoses. The individual comorbidities used in models are defined by methods described in van Walraven's implementation of the Elixhauser comorbidity score[18]. We adjust for 38 primary diagnoses drawn from the Healthcare Cost and Utilization (HCUP) Clinical Classification Software (CCS)[19], which consist of the top 20 most frequent single-level CCS diagnoses and 18 level-one multi-level categories of diagnoses (Appendix A.) In secondary

**BMJ** Open

analyses, to emphasize the role of data cleanliness, we estimate risk using *only* the laboratory values since the nonlaboratory values do not vary in data cleanliness and curation.

#### **Outcome Variable: 30-day mortality**

Our primary outcome variable is 30-day all-cause mortality, defined as death within 30 days of the admission date for the index hospitalization. Mortality is evaluated using the highly reliable Veterans Administration beneficiary death files which aggregate from multiple sources[12, 20, 21].

#### Statistical Analysis and Model Development

Random Forests is an ensemble machine learning method that aggregates the results of multiple decision trees fit on bootstrap samples of the original data[22, 23]. For each decision tree, the original data are bootstrapped to create a new dataset of the same size and the tree is fit to the new data. Instead of considering all predictors to determine the splitting criterion at a node, the split variable is chosen from a random subset of variables in order to reduce the correlation between different trees. Many such trees are grown, creating a 'forest'. Each observation is classified by each tree, and the majority classification over all trees is the predicted class. The ability of random forests to learn nonlinear and complex functions contributes to its predictive performance.

The neural network[24] can "learn" to classify samples without manual designed task-specific rules. The algorithm applies different weights to predictors and uses these transformations in subsequent "layers" of the neural net, culminating in the output layer with predictions. We applied the random forest and the neural network on our task. A traditional logistic regression model was also performed and compared.

Statistical analyses were performed with Python and the scikit-learn package[25].

Training and Testing Sets

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The dataset was randomly split into a 70% training set and a 30% testing set. The same split was used for all classification methods. This process was replicated five times (five different training sets and corresponding testing set were generated),

and each time the models were fit on the training set and used to predict the 30-day mortality of the testing set.

#### **Missing Data and Imputation**

We imputed the missing values before training and testing the models, comparing:

- "Mean Value": the mean value of each variable in the training set was used to replace missing values[26].
- "Random Forest": used random forest to impute missing values (missForest)[27].
- "Extremely Randomized Trees (Extra-Trees Regression)": this method is similar to random forest but is faster[28, 29].
- "Ridge Regression": used Bayesian Ridge regression to impute missing values[30].
- "Normal Value"[31]: normal values were used to impute missing values—this is common in clinical prediction contexts in which it is assumed that clinicians order tests they fear are not normal, and therefore the absence of such a test is a sign that the clinician reviewed other aspects of the patient's case and judged the odds of physiologic abnormality so low that testing was not indicated.
- "No Missing": case-wise deletion[32].

#### Variable Importance and Partial Dependence Plots

Predictor variable importance was evaluated for random forests[33]. When classifying a sample using a decision tree, a predictor was used at each node. Predictors that appear more frequently and that reduce the misclassification more substantially are considered more important. By combining all trees in a random forest model, we assessed the variable importance of each predictor. Different values of the same predictor may have different effects on the prediction. We plotted the Partial Dependence Plots[30] to show how the value of predictors affects the prediction of 30-day mortality. Partial dependence plots were used to visualize non-linearity among variables.

The cohort comprised 355,823 ICU hospitalizations at over 100 different hospitals, as described elsewhere[15]. The mean

age of the cohort was 66.9 years, and there were 34,867 deaths within 30-days of admission, a primary outcome event

**ICU Only Cohort** 

355,823

66.9 (11.6)

341,579 (96.0)

256,293 (72.0)

73,855 (20.8)

25,675 (7.2)

20,532 (5.8)

34,867 (9.8)

9.5 (13.0)

#### RESULTS

**Cohort Description** 

rate of 9.8% (Table 1.)

Hospitalizations, N

Age, mean (SD), y

Black or African American

30-day Mortality, N (%)

Length of Stay, mean (SD), days

Variables

Male, N (%)

Race, N (%)

Hispanic, N (%)

White

Other

 Table 1. ICU Patient Demographics

#### 

Rates of data missingness for each laboratory value in each dataset are shown in Table 2. Dataset A has a high proportion of missing laboratory values for blood urea nitrogen (0.84) and hematocrit (0.85) compared to datasets B and C. This is due to dataset A using a single, broad lab test name to identify laboratory values: "BUN" for blood urea nitrogen and "hematocrit" for hematocrit. In contrast, datasets B and C incorporated LOINC codes for BUN and HCT, which result in

fewer missing laboratory values.

### Table 2. Proportion of Labs Missing

Using all Data for Model Development

Dataset	Albumin (albval)	Bilirubin (bili)	Blood urea nitrogen (bun)	a Creatnine (creat)	Glucose (glucose)	Hematocrit (hct)	Partial Presssure (pao2)	pH (pa)	Sodium (na)	White Blood Cell (wbc)
A	0.39	0.42	0.84	0.13	0.07	0.85	0.66	0.14	0.11	0.13
в	0.38	0.42	0.13	0.13	0.06	0.12	0.65	0.44	0.11	0.13
с	0.39	0.45	0.13	0.12	0.06	0.11	0.69	0.64	0.11	0.13

Figure 1 shows the AUC scores of different classification models and imputation methods in the primary analysis. The highest AUC's obtained for each primary classification method (rows of the figure: logistic regression, random forest, or a neural network) were very similar: AUC's of 0.83 to 0.85. Likewise, there was relatively little variation within classification method by the missing value imputation method used, be it mean value imputation, random forest, extremely randomized trees (extra-trees regression), ridge regression, or normal value imputation. All models suffered dramatic losses in discrimination when case-wise deletion was used for missing data in the least clean dataset (far right columns). Full model performance for each condition can be seen in Appendix B.

Variation in discrimination was seen, however, across classification methods, as a function of data cleanliness. (Note that the analyst was blinded during the analysis to how each dataset was developed, and hence did not know which was "cleanest"). In the logistic regression model developed using the least clean data (dataset A had raw lab values extracted using only lab test names), performance was always lower than the performance with the more complete and clean datasets—by AUC's of 0.05 to about 0.1, p-value < 0.05). Similarly, performance in dataset B (extracted using LOINC codes without unit standardization) was lower and more unstable for mean value imputation and ridge regression. In marked contrast, neither random forests nor neural networks showed such reduced performance when developed in less clean data—in no case did the AUC degradations exceed 0.025 despite similar optimal performance.

#### Secondary Analysis Using only Laboratory Values

The primary analysis presented above considers the real world case in which demographics, diagnoses, and laboratory values are used in combination with risk model prediction. Yet, of these, only laboratory values were subject to variation in cleanliness. We, therefore, conducted a secondary analysis using only laboratory values to assess more clearly the impact of data quality. Results are shown in Figure 2.

Average model performance with this much smaller group of predictors is, as expected, somewhat lower with less data optimal AUC's typically range from 0.73 to 0.78 across combinations of classification model and missing data imputation. No uniformly superior strategy is evident, save markedly lower performance of case-wise deletion in the least clean

#### **BMJ** Open

dataset (A). As before, logistic regression shows markedly reduced discrimination when developed in the least clean data set. Neural networks show consistent performance.

Also notable is the marked reduction of discrimination of random forest models and neural network models regardless of the missing data imputation model used within dataset D. Dataset D has the "cleanest" data, in that it has hand-curated inclusion criteria, standardization of units, and conversion of values from their continuous scale to a semi-quantitative set of "points" as is done in the APACHE scoring algorithms. Attempting to work with such standardized point values as inputs consistently resulted in markedly worse discrimination in random forest models and neural network models than using other "less clean" datasets (the difference between Dataset D and other datasets is significant with a p-value < 0.05).

#### Variable Importance

The most important predictors of 30-day mortality were age and laboratory values. Age had the highest importance scores, regardless of which dataset was used, indicating that age is the most important variable when predicting 30-day mortality. The 10 laboratory values also had high importance scores. For datasets A, B, and C, laboratory values fell in the top-13 most important variables, and there were at least 8 laboratory values in the top-10 most important variables. However, for dataset D, there were only 6 laboratory values in the top-10 most important variables, and the variable white blood cell score ranked 20<sup>th</sup>. This may indicate that transforming laboratory values to APACHE scores results in the loss of information contained in the original values and negatively influences the performance of the random forest model.

#### **Partial Dependence Plots**

As it is hard to visualize the relationship between multiple predictors and the outcome, we created partial dependence plots to show the effect of predictors on the outcome [34]. The plots can also show whether the relationship between a specific predictor and the outcome is linear, guadratic, monotonic, or more complex. Further analysis can be done by combining the partial dependence plots and medical knowledge. Figure 3 and Figure 4 are the partial dependence plots for the pH score and the  $PaO_2$  score. We will take these as examples to show how the value of predictors in different datasets affects 30-day mortality. The X-axis is the value of the predictor. For each value of the predictor, the Y-axis is the averaged model output for all observations with the corresponding value of the predictor. A higher partial dependence

value corresponds to a higher risk of mortality. As we know, the normal value of the pH score is 7.4, and both higher values and lower values are abnormal. Typically, abnormal values lead to a larger risk of death. Therefore, a U-shaped partial dependence plot is to be expected for datasets A, B, and C. However, only the plot for dataset C is U-shaped. This is because dataset C is "cleaner" than datasets A and B, and the models can learn the real effect of pH score on 30-day mortality. Datasets A and B are not as clean as dataset C, as some other variables are presented in these datasets as pH score. Thus, it is difficult for the models to utilize the pH score variable in datasets A and B. This result indicates that cleaner variables benefits the classification models. However, not all variables have this problem. For most other variables such as the PaO<sub>2</sub> score, the plots of datasets A, B, and C have similar trends.

#### DISCUSSION

We used real data from a large nationwide health system to explore the interaction between missing data imputation techniques, data cleanliness, and classification methods for the common problem of predicting 30-day mortality in a holdout testing dataset. In brief, we found that any of several imputation techniques other than case-wise deletion performed equivalently in terms of discrimination, regardless of data cleanliness or classification method used. We found that logistic regression showed worse discrimination with less carefully cleaned data than did random forest or neural networks. Random forest models (and to a degree, neural networks) displayed diminished discrimination when given data that had been too highly cleaned and standardized prior to use.

#### Relationship to Past Research

Missing data are ubiquitous in large datasets. Even when missingness is completely at random, missing data lead to significant loss in statistical power and predictive ability[32]. We have previously found that the Random Forest method consistently produced the lowest imputation error compared to commonly used imputation methods[26]. Random Forest had the smallest prediction difference when 10-30% of the laboratory data was missing. Our present analysis of real data shows that as more specialized laboratory values are introduced into the prediction setting, much higher levels of missingness may be present. We thereby extend the previous finding that Random Forest continues to perform well for

#### **BMJ** Open

missing data. Our findings on the poor performance of case-wise deletion as an approach to handling missing data are in agreement with mainstream recommendations for more than two decades[32].

Our findings on missing data are of note because of the distinctive, yet real world, way in which missing data were generated. There were two missingness processes. First, clinicians in routine practice only sometimes order any given laboratory, and thus the presence or absence of an order may itself provide prognostic importance. [35] Second, a given effort to identify all of a given target laboratory values may or may not succeed. Even in a large system with a strong tradition of centralization, the extent to which laboratory ascension and labeling practices coincide with their aspiration varies over time, and often clinical insight is necessary to distinguish valid laboratory tests[36]. For any given data pull, it is not trivial to understand which missing values represent failure to find data that exist versus representing true missingness. Past work has rarely explicitly considered these distinct missingness-generating processes (in addition to true missingness at random) at their distinct implications.

The finding of poorer discrimination of Random Forest in models where the data were fully standardized and cleaned was not anticipated given past literature. The APACHE score was designed to simplify the lab results and to help doctors predict mortality [2]. Even in its more recent incarnations, APACHE transforms continuous lab results into discrete acute physiology scores[37]. Our data suggest that transforming lab results to APACHE scores is not necessary for Random Forest and may even lead to the loss of information[23]. Remarkably, even standardization to equivalent units across institutions may not be necessary—but at the same time, this means that sources of variance other than simply the laboratory value may also be subtly incorporated into risk-prediction with non-standardized ways. It is a case-specific decision as to whether incorporation of such variance is helpful for a given task or is a source of bias.

#### Implications

Our findings have implications for both practitioners seeking to implement a given prediction rule and scientists interested in risk-prediction generally. For practitioners, no given method yields consistently superior results in terms of discrimination. Therefore, other performance considerations, whether psychometric or implementation ease, may play For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

an important role. They also suggest that missing data imputation approaches other than case-wise deletion during development are mandatory.

Our results also note that Random Forests and neural networks were strikingly robust to even quite naively prepared data, in contrast to logistic regression. This suggests that the truth of the oft-quoted aphorisms about "garbage in, garbage out" may depend on the categorization model and missing data imputation method used. In situations where ascertainment and cleaning of data are more costly, random forests may offer pragmatic advantages if these findings are replicable.

#### Strengths and Limitations

Strengths of our analysis include its use of real world data, with real world data generation and missingness-generation problems on an established problem encountered by medical researchers and clinicians. We also used multiple methods implemented in standardized ways. The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility.

Limitations of our analysis stem fundamentally from the nearly infinite combinations of analysis factors that might be varied, and our inability to explore such a high dimensional space. Thus we only considered one outcome and one standardization method, and decided upon an a priori approach for each combination of dataset, categorization model, and missingness imputation method used. Other outcomes and other possible data structures (such as using trends in data) may yield different answers. We focus on discrimination, as measured by AUC, but other measurement properties are assuredly also important. We also focused on individual-level prediction, as opposed to considering the impact on hospital-level quality assessment or other tasks for which these results may be used.

#### CONCLUSION

In sum, our results suggest that there is little variation in discrimination among different statistical classification models in well-cleaned data using modern missing data imputation techniques. As such, the decision about which of the well-

#### Page 15 of 34

#### BMJ Open

1	performing imputation and adjustment methods to use can be made based on other factors relevant to the particular
2 3	application—as long as the lower performing methods are avoided. If these findings are replicated in other data with other
54 55	
56 57	
58 59	14
60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### **Figure Captions**

- Figure 1. AUC Scores, Full Model
- Figure 2. AUC Scores for lab-only predictors
- Figure 3. Partial Dependence Plots for pH
- Figure 4. Partial Dependence Plots for PaO2

torbeer terien only

#### ACKNOWLEDGEMENTS

#### Author Contributions

**Theodore J. Iwashyna:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing

Cheng Ma: Formal analysis, software, visualization, writing-original draft, and writing-review & editing

Xiao Qing Wang: Data curation; Writing – original draft; Writing – review & editing

Sarah Seelye : Writing – original draft; Writing – review & editing

Ji Zhu : Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

Akbar K. Waljee: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

**Declarations of funding interests:** This work was supported by a career development grant award [CDA 11-217 to A.K.W] and a Merit Review Award Number [IIR 16-024 to A.K.W, 17-045 to T.J.I.] from the United States Department of Veterans Affairs Health Services Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Michigan, the Veterans Affairs, the U.S. Government, or the National Institutes of Health.

Disclosures: The authors declare there are no conflicts to disclose

**Guarantor of the article:** Akbar K. Waljee the submission's guarantor, takes responsibility for the integrity of the work as a whole, from inception to published article.

#### Licensing Statement:

I, Akbar Waljee, as the Submitting Author have the right to grant and does grant on behalf of all authors of the Work (as defined in the below author license), an exclusive license and/or a non-exclusive license for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY license shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our <u>license</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons license – details of these licenses and which <u>Creative Commons</u> license will apply to this Work are set out in our license referred to above.

## 53 Data Sharing Statement:54

Appendices and statistical code are available via Github at <a href="https://github.com/CCMRcodes/GIVO">https://github.com/CCMRcodes/GIVO</a>. The dataset cannot be
 disseminated due to inclusion of sensitive patient information under VA regulations.

 tor peer terier only

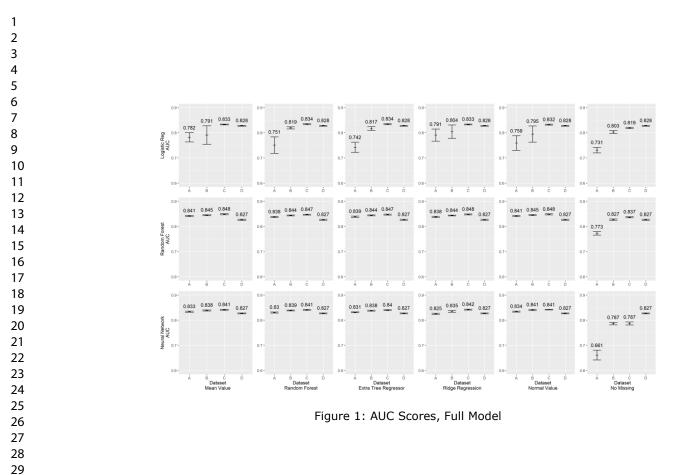
#### REFERENCES

- [1] Iezzoni LI. Risk adjustment for measuring health care outcomes. 4th ed. Chicago, Ill. Arlington, VA: Health
   Administration Press; AUPHA; 2013.
- [2] Lane-Fall MB, Neuman MD. Outcomes measures and risk adjustment. Int Anesthesiol Clin. 2013;51:10-21.
- [2] Lane-Fairwis, Neuman WD. Outcomes measures and risk adjustment. Int Anestnesion (
   [3] Quality AfHRa. Part II. Introduction to Measures of Quality (continued). Rockville, MD.
- [4] Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, et al. Efficacy and safety of
- 8 recombinant human activated protein C for severe sepsis. N Engl J Med. 2001;344:699-709.
- [5] Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction.
  Stat Med. 1984;3:143-52.
- [6] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions
   and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
- [7] van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10
   events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163.
- [8] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a
- <sup>17</sup> multivariable prediction model: PART II binary and time-to-event outcomes. Stat Med. 2019;38:1276-96.
- [9] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic
   prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28:2455-74.
- [10] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a
- multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern
   Med. 2015;162:W1-73.
   Med. 2015;162:W1-73.
- [11] Steyerberg EW, ProQuest (Firm). Clinical prediction models a practical approach to development, validation, and
   updating. New York: Springer; 2009. p. xxviii, 497 p.
- [12] Render ML, Kim HM, Welsh DE, Timmons S, Johnston J, Hui S, et al. Automated intensive care unit risk adjustment: results from a National Veterans Affairs study. Crit Care Med. 2003;31:1638-46.
- [13] Render ML, Deddens J, Freyberg R, Almenoff P, Connors AF, Wagner D, et al. Veterans Affairs intensive care unit risk
   adjustment model: validation, updating, recalibration. Crit Care Med. 2008;36:1031-42.
- <sup>30</sup> [14] Render ML, Freyberg RW, Hasselbeck R, Hofer TP, Sales AE, Deddens J, et al. Infrastructure for quality
- transformation: measurement and reporting in veterans administration intensive care units. BMJ Qual Saf. 2011;20:498-507.
- [15] Wang XQ, Vincent BM, Wiitala WL, Luginbill KA, Viglianti EM, Prescott HC, et al. Veterans Affairs patient database (VAPD 2014-2017): building nationwide granular data for clinical discovery. DMC Model Dev Market Dev Market Dev
- (VAPD 2014-2017): building nationwide granular data for clinical discovery. BMC Med Res Methodol. 2019;19:94.
- [16] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying
   laboratory observations: a 5-year update. Clin Chem. 2003;49:624-33.
- 38 [17] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and
- codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results.
   Clin Chem. 1996;42:81-90.
- [18] van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures
   inter a neint system for hearing a derivative state. Mad. Comp. 2000; 17:222-22
- into a point system for hospital death using administrative data. Med Care. 2009;47:626-33.
   [40] UCUP US Tools & Software Page 2010.
- [19] HCUP-US Tools & Software Page. 2019.
- [20] Hooper TI, Gackstetter GD, Leardmann CA, Boyko EJ, Pearse LA, Smith B, et al. Early mortality experience in a large
   military cohort and a comparison of mortality data sources. Popul Health Metr. 2010;8:15.
- [21] Prescott HC, Kepreos KM, Wiitala WL, Iwashyna TJ. Temporal Changes in the Influence of Hospitals and Regional
- 48 Healthcare Networks on Severe Sepsis Mortality. Crit Care Med. 2015;43:1368-74.
- 49 [22] Breiman L. Classification and regression trees. New York, NY: Chapman & Hall; 1993.
- [23] Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
   [24] Omiduar O. Dayboff JE. Science Direct (Online convice). Neural
- [24] Omidvar O, Dayhoff JE, ScienceDirect (Online service). Neural networks and pattern recognition. San Diego, Calif.:
   Academic Press; 1998.
   [25] Dedregees E. Veregeugung C. Complete A. Michael V. Thinkson D. Calif. I. Status and San Diego. Calif.
- [25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python.
   Journal of machine learning research. 2011;12:2825–30.
- 56
- 57
- 58 59

[26] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3.

- [27] Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data.
- Bioinformatics. 2012;28:112-8.

- [28] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63:3-42.
- [29] Maree R, Geurts P, Wehenkel L. Random subwindows and extremely randomized trees for image classification in cell biology. BMC Cell Biol. 2007;8 Suppl 1:S2.
- [30] Hastie T, Friedman J, Tibshirani R, SpringerLink (Online service). The Elements of Statistical Learning Data Mining, Inference, and Prediction. New York, NY: Springer New York : Imprint: Springer; 2001.
- [31] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health
- evaluation: a physiologically based classification system. Crit Care Med. 1981;9:591-7.
- [32] Allison PD, Sage Publications. Missing data. Thousand Oaks, [Calif.] ; London: SAGE; 2002. p. 1 online resource (vi, 91 p.).
- [33] Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees.
- Proceedings of the 26th International Conference on Neural Information Processing Systems Volume 1. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. p. 431-9.
- [34] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001;29:1189-232.
- [35] Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.
- [36] Wiitala WL, Vincent BM, Burns JA, Prescott HC, Waljee AK, Cohen GR, et al. Variation in Laboratory Test Naming
- Conventions in EHRs Within and Between Hospitals: A Nationwide Longitudinal Study. Med Care. 2019;57:e22-e7.
- [37] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV:
- hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34:1297-310.



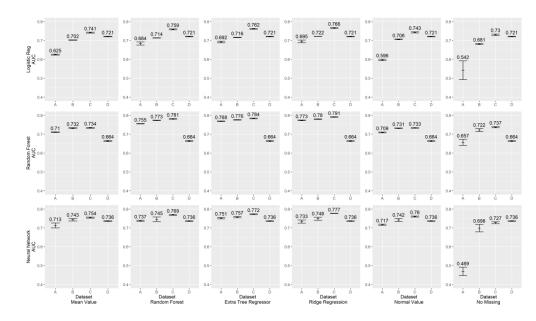
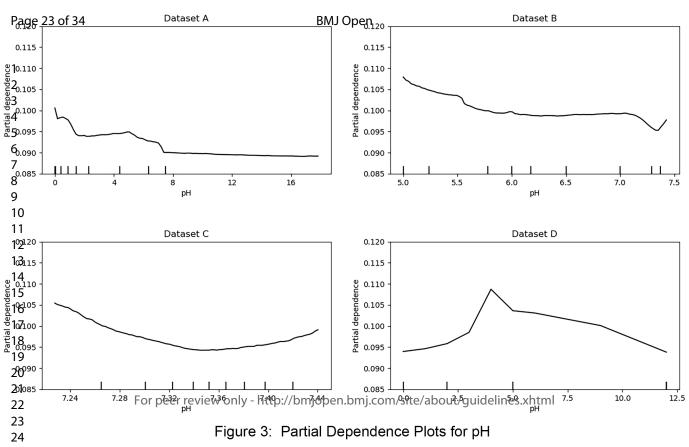
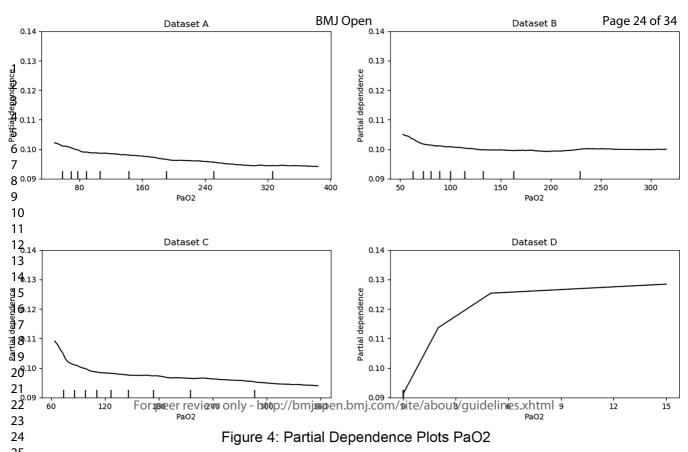


Figure 2. AUC Scores for lab-only predictors





#### Appendix A. Patient-level variables included in models

Demographics	Gender, Age, Race (White, Black or African American, Asian, Native
	Hawaiian or other Pacific Islander, Unknown), Hispanic ethnicity
	Hypertension, Congestive Heart Failure, Cardiac Arrhythmia, Valvular
	Disease, Pulmonary Circulation Disorders, Peripheral Vascular
	Disorders, Paralysis, Other Neurological Disorders, Chronic Pulmonary
Comorbidities,	Disease, Diabetes Uncomplicated, Diabetes Complicated,
included in Elixhauser	Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer Disease
	excluding bleeding, AIDS/HIV, Lymphoma, Metastatic Cancer, Solid
	Tumor without Metastasis, Rheumatoid Arthritis/Collagen,
	Coagulopathy, Obesity, Weight Loss, Fluid and Electrolyte Disorders,
	Blood Loss Anemia, Deficiency Anemia, Alcohol Abuse, Drug Abuse,
	Psychoses, Depression
	Top 20 most frequent single-level CCS diagnoses:
	Congestive Heart Failure (non-hypertensive), Non-specific Chest Pain,
	Coronary Atherosclerosis and Other Heart Disease, Cardiac
	Dysrhythmias, Alcohol-related Disorders, Septicemia (except in labor),
	Chronic Obstructive Pulmonary Disease and Bronchiectasis,
	Pneumonia, Skin and Subcutaneous Tissue Infections, Osteoarthritis,
	Complication of Device (implant or graft), Complications of Surgical
	Procedures or Medical Care, Diabetes Mellitus with Complications,
	Respiratory Failure, Urinary Tract Infections, Renal Failure,
Diagnoses,	Spondylosis, Acute Myocardial Infarction, Fluid and Electrolyte
HCUP CCS single-level and multi-level	Disorders, Gastrointestinal Hemorrhage
	bisorders, edistrointestinar nemorninge
	18 level 1 multi-level CCS categories:
	Infectious and Parasitic Diseases, Neoplasms, Endocrine Disorders,
	Anemia, Mental Illness, Diseases of the Nervous System, Diseases of
	the Circulatory System, Diseases of the Respiratory System, Diseases of
	the Digestive System, Diseases of the Genitourinary System,
	Complications of Pregnancy or Childbirth, Skin Disease, Diseases of the
	Musculoskeletal System, Congenital Anomalies, Perinatal Conditions, Injury and Poisoning, Other Health Status Conditions, Other Residual
	Codes
	Albumin, Bilirubin, Blood Urea Nitrogen, Creatinine, Glucose,
	Abamin, binabin, bioda orea Micogen, creatinine, diacose,
Laboratory values	Hematocrit, Partial pressure of oxygen score, pH score, Sodium, White

#### Appendix B

#### Table B.1: Model Performances (Full Model)

Classification		Imputation	AUROC	Optimal	Predicted Cases		ases Accurate Rate				Brier
Method	Dataset	Method	(95%CI)	Cutoff	Death	Survival	Death	Survival	Sensitivity	Specificity	Score
Logistic Regression	A	Mean Value	0.78(0.76- 0.80)	0.48	37199	69549	0.21	0.96	0.74	0.69	0.19
Logistic Regression	В	Mean Value	0.79(0.75- 0.83)	0.46	34915	71833	0.24	0.97	0.80	0.72	0.17
Logistic Regression	С	Mean Value	0.83(0.83- 0.83)	0.46	35970	70778	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Random Forest	0.75(0.72- 0.78)	0.49	39528	67220	0.18	0.95	0.69	0.66	0.21
Logistic Regression	В	Random Forest	0.82(0.82- 0.82)	0.49	31996	74752	0.25	0.97	0.77	0.75	0.17
Logistic Regression	С	Random Forest	0.83(0.83- 0.84)	0.46	36017	70731	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Extra Trees Regression	0.74(0.72- 0.76)	0.46	45762	60986	0.17	0.96	0.74	0.61	0.21
Logistic Regression	В	Extra Trees Regression	0.82(0.81- 0.82)	0.47	33642	73106	0.25	0.97	0.79	0.74	0.17
Logistic Regression	С	Extra Trees Regression	0.83(0.83- 0.84)	0.47	34579	72169	0.24	0.97	0.78	0.73	0.17
Logistic Regression	A	Ridge Regression	0.79(0.77- 0.82)	0.46	38579	68169	0.21	0.97	0.78	0.68	0.18
Logistic Regression	В	Ridge Regression	0.80(0.78- 0.83)	0.46	35034	71714	0.24	0.97	0.80	0.72	0.17

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	27	of	34
------	----	----	----

		<b>D</b> : 1	0.00/0.00								
Logistic Regression	С	Ridge Regression	0.83(0.83- 0.84)	0.47	35220	71528	0.23	0.97	0.78	0.72	0.17
Logistic Regression	A	Normal Value	0.76(0.73- 0.79)	0.50	37392	69356	0.18	0.95	0.63	0.68	0.22
Logistic Regression	В	Normal Value	0.80(0.76- 0.83)	0.50	30977	75771	0.26	0.97	0.76	0.76	0.17
Logistic Regression	С	Normal Value	0.83(0.83- 0.83)	0.46	35676	71072	0.23	0.97	0.78	0.72	0.17
Logistic Regression	A	No Missing	0.73(0.72- 0.74)	0.43	37658	69090	0.18	0.95	0.66	0.68	0.18
Logistic Regression	В	No Missing	0.8(0.80- 0.81)	0.42	33153	73595	0.24	0.97	0.76	0.74	0.14
Logistic Regression	С	No Missing	0.82(0.82- 0.82)	0.44	35333	71415	0.22	0.96	0.76	0.72	0.16
Logistic Regression	D	None	0.83(0.83- 0.83)	0.47	34184	72564	0.24	0.97	0.78	0.73	0.17
Random Forest	A	Mean Value	0.84(0.84- 0.84)	0.12	32330	74418	0.26	0.97	0.79	0.75	0.07
Random Forest	В	Mean Value	0.85(0.84- 0.85)	0.11	32642	74106	0.26	0.97	0.80	0.75	0.07
Random Forest	С	Mean Value	0.85(0.85- 0.85)	0.11	33548	73200	0.25	0.97	0.81	0.74	0.07
Random Forest	A	Random Forest	0.84(0.84- 0.84)	0.12	32659	74089	0.25	0.97	0.78	0.75	0.07
Random Forest	В	Random Forest	0.84(0.84- 0.85)	0.11	34093	72655	0.25	0.97	0.81	0.73	0.07

**BMJ** Open

Random		Random	0.85(0.85-	0.11	33029	73719	0.25	0.97	0.80	0.74	0.0
Forest	С	Forest	0.85)	0.11	55029	/3/19	0.25	0.97	0.80	0.74	0.0
Random		Extra Trees	0.84(0.84-	0.11	32938	73810	0.25	0.97	0.79	0.74	0.0
Forest	А	Regression	0.84)	0.11	32938	/3810	0.25	0.97	0.79	0.74	0.0
Random		Extra Trees	0.84(0.84-	0.12	22411	74227	0.26	0.07	0.80	0.75	0.0
Forest	В	Regression	0.85)	0.12	32411	74337	0.26	0.97	0.80	0.75	0.0
Random		Extra Trees	0.85(0.85-	0.11	33567	73181	0.25	0.97	0.80	0.74	0.0
Forest	С	Regression	0.85)	0.11	55507	/5101	0.25	0.97	0.80	0.74	0.0
Random		Ridge	0.84(0.45-	0.11	34587	72161	0.24	0.97	0.80	0.73	0.0
Forest	A	Regression	0.84)		54507	/2101	0.24	0.57	0.00	0.75	0.0
Random		Ridge	0.84(0.84-	0.12	31643	75105	0.26	0.97	0.79	0.76	0.0
Forest	В	Regression	0.85)	0.12	51045	/ 5105	0.20	0.57	0.75	0.70	0.0
Random		Ridge	0.85(0.85-	0.12	32531	74217	0.25	0.97	0.79	0.75	0.0
Forest	С	Regression	0.85)	0.12	52551		0.23	0.57	0.75	0.75	0.0
Random		Normal	0.84(0.84-	0.12	31234	75514	0.26	0.97	0.78	0.76	0.0
Forest	A	Value	0.84)	0.12	51254	75514	0.20	0.57	0.70	0.70	0.0
Random		Normal	0.85(0.84-	0.11	32711	74037	0.26	0.97	0.80	0.75	0.0
Forest	В	Value	0.85)	0.11	52711	74037	0.20	0.57	0.00	0.75	0.0
Random		Normal	0.85(0.85-	0.12	31159	75589	0.26	0.97	0.78	0.76	0.0
Forest	С	Value	0.85)	0.12	51155	75505	0.20	0.57	0.70	0.70	0.0
Random			0.77(0.77-	0.18	36332	70416	0.20	0.96	0.71	0.70	0.0
Forest	А	No Missing	0.78)	0.10	50552	/0410	0.20	0.90	0.71	0.70	0.0
Random			0.83(0.82-	0.16	31836	74912	0.25	0.97	0.77	0.75	0.0
Forest	В	No Missing	0.83)	0.10	21020	/4J12	0.25	0.57	0.77	0.75	0.0
Random			0.84(0.83-	0.16	34517	72231	0.24	0.97	0.78	0.73	0.0
Forest	С	No Missing	0.84)	0.10	5451/	12231	0.24	0.97	0.76	0.75	0.0

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 29 of 34

B	MJ Open	

Random			0.83(0.83-	0.11	33407	73341	0.24	0.97	0.77	0.74	
Forest	D	None	0.83)	0.11	00107	/0011	0.21	0.07	0177	0.71	
Neural		Mean	0.83(0.83-	0.53	35031	71717	0.24	0.97	0.80	0.72	
Network	А	Value	0.84)	0.55	55051	, 1, 1,	0.24	0.57	0.00	0.72	
Neural		Mean	0.84(0.84-	0.52	32716	74032	0.25	0.97	0.79	0.75	
Network	В	Value	0.84)	0.52	52710	74032	0.25	0.57	0.75	0.75	
Neural		Mean	0.84(0.84-	0.53	32549	74199	0.25	0.97	0.79	0.75	
Network	С	Value	0.84)	0.55	52549	74199	0.25	0.97	0.79	0.75	
Neural		Random	0.83(0.83-	0.55	32515	74233	0.25	0.97	0.77	0.75	
Network	А	Forest	0.83)	0.55	52515	74255	0.25	0.97	0.77	0.75	
Neural		Random	0.84(0.84-	0.57	30842	75906	0.26	0.97	0.77	0.76	
Network	В	Forest	0.84)	0.57	50042	75500	0.20	0.57	0.77	0.70	
Neural		Random	0.84(0.84-	0.55	34144	72604	0.24	0.97	0.79	0.73	
Network	С	Forest	0.84)	0.55	54144	72004	0.24	0.57	0.75	0.75	
Neural		Extra Trees	0.83(0.83-	0.50	37351	69397	0.23	0.97	0.82	0.70	
Network	А	Regression	0.83)	0.50	57551	05557	0.25	0.57	0.02	0.70	
Neural		Extra Trees	0.84(0.84-	0.54	33529	73219	0.25	0.97	0.80	0.74	
Network	В	Regression	0.84)	0.54	33329	73219	0.25	0.97	0.80	0.74	
Neural		Extra Trees	0.84(0.84-	0.49	33324	73424	0.25	0.97	0.79	0.74	
Network	С	Regression	0.84)	0.45	55524	73424	0.25	0.57	0.75	0.74	
Neural		Ridge	0.83(0.82-	0.51	33864	72884	0.24	0.97	0.78	0.73	
Network	А	Regression	0.83)	0.51	55004	/2004	0.24	0.97	0.70	0.75	
Neural		Ridge	0.84(0.83-	0.52	31186	75562	0.26	0.97	0.78	0.76	(
Network	В	Regression	0.84)	0.52	21190	10002	0.20	0.97	0.78	0.70	
Neural		Ridge	0.84(0.84-	0.55	22145	74602	0.25	0.07	0.79	0.75	
Network	С	Regression	0.84)	0.55	32145	74603	0.25	0.97	0.78	0.75	

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	30 of	34
------	-------	----

									0 1		
Network	D	None	0.83)			Q	10				
Neural		Nono	0.83(0.83-	0.52	33760	72988	0.24	0.97	0.78	0.73	0.18
Neural Network	С	No Missing	0.79(0.78- 0.79)	0.59	38619	68129	0.20	0.96	0.75	0.68	0.24
Neural Network	В	No Missing	0.79(0.78- 0.79)	0.59	32676	74072	0.23	0.96	0.71	0.74	0.21
Neural Network	A	No Missing	0.66(0.64- 0.68)	0.76	49011	57737	0.14	0.94	0.65	0.56	0.50
Neural Network	С	Normal Value	0.84(0.84- 0.84)	0.55	32864	73884	0.25	0.97	0.79	0.75	0.18
Neural Network	В	Normal Value	0.84(0.84- 0.84)	0.45	35026	71722	0.24	0.97	0.81	0.72	0.16
Neural Network	А	Normal Value	0.83(0.83- 0.84)	0.58	31675	75073	0.25	0.97	0.77	0.76	0.19

Classification		Imputation	AUROC	Optimal	Predict	ed Cases	Accura	ite Rate			Brie
Method	Dataset	Method	(95%CI)	Cutoff	Death	Survival	Death	Survival	Sensitivity	Specificity	Sco
Logistic Regression	A	Mean Value	0.63(0.62- 0.63)	0.50	32327	74421	0.16	0.93	0.50	0.72	0.24
Logistic Regression	В	Mean Value	0.70(0.70- 0.70)	0.47	33350	73398	0.21	0.95	0.66	0.73	0.20
Logistic Regression	С	Mean Value	0.74(0.74- 0.74)	0.47	35248	71500	0.19	0.95	0.62	0.70	0.22
Logistic Regression	A	Random Forest	0.68(0.68- 0.69)	0.48	39566	67182	0.17	0.94	0.63	0.66	0.23
Logistic Regression	В	Random Forest	0.71(0.71- 0.71)	0.45	37758	68990	0.20	0.96	0.71	0.69	0.20
Logistic Regression	С	Random Forest	0.76(0.76- 0.76)	0.47	38421	68327	0.18	0.95	0.66	0.67	0.21
Logistic Regression	A	Extra Trees Regression	0.69(0.69- 0.70)	0.46	43607	63141	0.17	0.95	0.69	0.62	0.22
Logistic Regression	В	Extra Trees Regression	0.72(0.72- 0.72)	0.44	39295	67453	0.20	0.96	0.73	0.67	0.19
Logistic Regression	С	Extra Trees Regression	0.76(0.76- 0.76)	0.45	42675	64073	0.17	0.95	0.71	0.63	0.21
Logistic Regression	A	Ridge Regression	0.70(0.69- 0.70)	0.47	42514	64234	0.17	0.95	0.69	0.63	0.22
Logistic Regression	В	Ridge Regression	0.72(0.72- 0.72)	0.44	39856	66892	0.20	0.96	0.75	0.67	0.19

Page	32	of	34
· ~	-	•••	•

Logistic		Ridge	0.77(0.76-	0.45	42737	64011	0.17	0.95	0.71	0.63	0.2
Regression	С	Regression	0.77)								
Logistic		Normal	0.60(0.59-	0.49	31990	74758	0.15	0.92	0.46	0.72	0.2
Regression	А	Value	0.60)	0.45	51550	/4/30	0.15	0.52	0.40	0.72	0.2
Logistic		Normal	0.71(0.70-	0.45	38325	68423	0.19	0.95	0.70	0.68	0.2
Regression	В	Value	0.71)	0.45	30323	00425	0.19	0.95	0.70	0.08	0.2
Logistic		Normal	0.74 (0.74-	0.46	41447	65301	0.17	0.95	0.68	0.64	0.2
Regression	С	Value	0.75)	0.40	41447	05301	0.17	0.95	0.08	0.64	0.2
Logistic			0.54 (0.49-	0.57	17678	89070	0.15	0.91	0.25	0.84	0.2
Regression	A	No Missing	0.59)		1,0,0	00070	0.15	0.51	0.20	0.04	0.2
Logistic			0.68 n(0.68-	0.45	32766	73982	0.20	0.95	0.64	0.73	0.1
Regression	В	No Missing	0.68)	0.45	32700	/ 5502	0.20	0.55	0.04	0.75	0.1
Logistic			0.73(0.73-	0.50	30965	75783	0.19	0.94	0.55	0.74	0.2
Regression	С	No Missing	0.73)	0.50	30303	/3/03	0.13	0.54	0.55	0.74	0.2
Logistic			0.72(0.72-	0.49	35766	70982	0.19	0.95	0.64	0.70	0.2
Regression	D	None	0.72)	0.45	55700	70502	0.15	0.55	0.04	0.70	0.2
Random		Mean	0.71(0.71-	0.09	46226	60522	0.17	0.95	0.73	0.60	0.0
Forest	А	Value	0.71)	0.09	40220	00322	0.17	0.95	0.75	0.00	0.0
Random		Mean	0.73(0.73-	0.10	44628	62120	0.18	0.96	0.75	0.62	0.0
Forest	В	Value	0.73)	0.10	44020	02120	0.10	0.50	0.75	0.02	0.0
Random		Mean	0.73(0.73-	0.10	44525	62223	0.18	0.96	0.75	0.62	0.0
Forest	С	Value	0.74)	0.10	44323	02223	0.10	0.90	0.75	0.02	0.0
Random		Random	0.76(0.75-	0.09	41715	65033	0.18	0.96	0.74	0.65	0.0
Forest	А	Forest	0.76)	0.09	41/13	05033	0.18	0.90	0.74	0.05	0.0
Random		Random	0.77(0.77-	0.10	38154	68594	0.20	0.96	0.75	0.69	0.0
Forest	В	Forest	0.77)	0.10	50154	00334	0.20	0.90	0.75	0.05	0.0

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 33	of 34
---------	-------

Random Forest	С	Random Forest	0.78(0.78- 0.78)	0.09	40709	66039	0.19	0.96	0.75	0.66	0.08
Random Forest	А	Extra Trees Regression	0.77(0.77- 0.77)	0.09	43230	63518	0.19	0.96	0.77	0.64	0.08
Random Forest	В	Extra Trees Regression	0.78(0.77- 0.78)	0.10	38734	68014	0.20	0.96	0.76	0.68	0.08
Random Forest	С	Extra Trees Regression	0.78(0.78- 0.79)	0.10	38810	67938	0.20	0.96	0.74	0.68	0.08
Random Forest	А	Ridge Regression	0.77(0.77- 0.78)	0.10	39913	66835	0.20	0.96	0.75	0.67	0.08
Random Forest	В	Ridge Regression	0.78(0.78- 0.78)	0.09	39663	67085	0.20	0.97	0.77	0.67	0.08
Random Forest	С	Ridge Regression	0.79(0.79- 0.79)	0.09	40249	66499	0.20	0.96	0.76	0.66	0.08
Random Forest	А	Normal Value	0.71(0.71- 0.71)	0.10	46047	60701	0.17	0.95	0.73	0.60	0.09
Random Forest	В	Normal Value	0.73(0.73- 0.73)	0.10	44400	62348	0.18	0.96	0.75	0.62	0.09
Random Forest	С	Normal Value	0.73(0.73- 0.74)	0.09	46774	59974	0.17	0.96	0.77	0.60	0.09
Random Forest	A	No Missing	0.66(0.64- 0.67)	0.26	20159	86589	0.21	0.93	0.40	0.83	0.10
Random Forest	В	No Missing	0.72(0.71- 0.73)	0.14	52201	54547	0.16	0.96	0.78	0.54	0.08
Random Forest	С	No Missing	0.74(0.73- 0.74)	0.21	26193	80555	0.22	0.94	0.55	0.79	0.09

Random			0.66(0.66-	0.45	33868	72880	0.18	0.94	0.57	0.71	0.18
Forest	D	None	0.67)	0.45	33000	72000	0.10	0.54	0.07	0.71	0.10
Neural		Mean	0.71(0.70-	0.47	33169	73579	0.19	0.94	0.60	0.72	0.19
Network	А	Value	0.73)	0.47	55105	13313	0.15	0.54	0.00	0.72	0.13
Neural		Mean	0.74(0.74-	0.49	31171	75577	0.21	0.95	0.63	0.75	0.19
Network	В	Value	0.75)	0.49	511/1	13377	0.21	0.95	0.05	0.75	0.15
Neural		Mean	0.75(0.75-	0.40	42096	64652	0.18	0.96	0.74	0.64	0.10
Network	С	Value	0.76)	0.40	42096	04052	0.18	0.96	0.74	0.64	0.18
Neural		Random	0.74(0.74-	0.50	37930	68818	0.19	0.95	0.68	0.68	0.19
Network	А	Forest	0.74)	0.50	37330	00010	0.15	0.55	0.00	0.00	0.11
Neural		Random	0.75(0.73-	0.54	36554	70194	0.20	0.96	0.71	0.70	0.22
Network	В	Forest	0.76)	0.54	50554	/0154	0.20	0.50	0.71	0.70	0.22
Neural		Random	0.77(0.77-	0.42	42444	64304	0.18	0.96	0.74	0.64	0.18
Network	С	Forest	0.77)	0.42	72777	04504	0.10	0.50	0.74	0.04	0.10
Neural		Extra Trees	0.75(0.75-	0.42	44585	62163	0.18	0.96	0.76	0.62	0.18
Network	А	Regression	0.76)	0.42	44303	02105	0.18	0.90	0.70	0.02	0.10
Neural		Extra Trees	0.76(0.75-	0.46	40067	66681	0.20	0.96	0.75	0.67	0.19
Network	В	Regression	0.76)	0.40	40007	00001	0.20	0.50	0.75	0.07	0.1.
Neural		Extra Trees	0.77(0.77-	0.48	39088	67660	0.19	0.96	0.71	0.67	0.20
Network	С	Regression	0.77)	0.40	55000	07000	0.15	0.50	0.71	0.07	0.20
Neural		Ridge	0.73(0.73-	0.49	43129	63619	0.18	0.96	0.74	0.63	0.22
Network	А	Regression	0.74)	0.73	73123	03013	0.10	0.90	0.74	0.05	0.2.
Neural		Ridge	0.75(0.74-	0.44	39388	67360	0.20	0.96	0.74	0.67	0.18
Network	В	Regression	0.76)	0.44	32300	0/500	0.20	0.90	0.74	0.07	0.10
Neural		Ridge	0.78(0.78-	0.53	38048	68700	0.19	0.95	0.68	0.68	0.22
Network	С	Regression	0.78)	0.55	50040	00700	0.15	0.55	0.00	0.00	0.21

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	35	of	34
------	----	----	----

 BMJ Open

Neural Network	А	Normal Value	0.72(0.71- 0.72)	0.45	33193	73555	0.19	0.95	0.61	0.72	0.18
Neural Network	В	Normal Value	0.74(0.73- 0.75)	0.50	37560	69188	0.20	0.96	0.71	0.69	0.20
Neural Network	С	Normal Value	0.76(0.76- 0.76)	0.41	42187	64561	0.18	0.96	0.74	0.64	0.18
Neural Network	А	No Missing	0.47(0.45- 0.49)	0.79	2628	104120	0.14	0.90	0.04	0.98	0.44
Neural Network	В	No Missing	0.70(0.68- 0.72)	0.69	23720	83028	0.23	0.94	0.52	0.81	0.31
Neural Network	С	No Missing	0.73(0.72- 0.73)	0.57	51077	55671	0.15	0.95	0.74	0.55	0.29
Neural Network	D	None	0.74(0.73- 0.74)	0.50	36925	69823	0.19	0.95	0.67	0.69	0.21
										0.69	

# **BMJ Open**

#### Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU Mortality, a US Nationwide Retrospective Cohort and Simulation Study

Journal:	BMJ Open
Manuscript ID	bmjopen-2020-041421.R2
Article Type:	Original research
Date Submitted by the Author:	04-Nov-2020
Complete List of Authors:	Iwashyna, Theodore; University of Michigan, Internal Medicine; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Ma, Cheng; University of Michigan, Statistics Wang, Xiao Qing; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Seelye, Sarah; VA Center for Clinical Management Research, Ann Arbor VA Medical Center Zhu, Ji; University of Michigan, Department of Statistics Waljee, Akbar; University of Michigan, Gastroenterology; VA Center for Clinical Management Research, Ann Arbor VA Medical Center
<b>Primary Subject Heading</b> :	Research methods
Secondary Subject Heading:	Health services research, Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, STATISTICS & RESEARCH METHODS, Adult intensive & critical care < INTENSIVE & CRITICAL CARE

SCHOLARONE<sup>™</sup> Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our <u>licence</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which <u>Creative Commons</u> licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

reliez oni

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Title: Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU ماريمينا والمربية المربية

1 2 3	Mortality, a US Nationwide Retrospective Cohort and Simulation Study
4 5	Authors:
6 7	Theodore J. Iwashyna*, M.D., PhD (1,2,3)
8 9	Cheng Ma*, B.S. (4)
10 11	Xiao Qing Wang, MPH (1)
12	Sarah Seelye, Ph.D. (1)
13 14	Ji Zhu, Ph.D. (4)
15 16	Akbar K. Waljee, M.D., M.Sc. (1,2,3)
17 18 19 20 21	<u>*considered co-first authors.</u>
22 23 24 25 26 27	<ol> <li>VA Center for Clinical Management Research, VA Ann Arbor Health Care System, Ann Arbor, MI, USA.</li> <li>Department of Internal Medicine, Michigan Medicine, Ann Arbor, MI, USA.</li> <li>Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP), Ann Arbor, MI, USA</li> <li>Department of Statistics, University of Michigan, Ann Arbor, MI, USA.</li> </ol>
28 29 30 31 32	Address correspondence to: Akbar K. Waljee, M.D. M.Sc., Associate Professor of Medicine, Division of Gastroenterology, Department of Internal Medicine, Michigan Medicine, Ann Arbor Veterans Affairs Medical Center, 2215 Fuller Road, 111D, Ann Arbor, Michigan 48105
33 34	Phone: 734-845-5865; Fax: 734-845-3091; E-mail: <u>awaljee@med.umich.edu</u>
35	
36 37 38	Keywords: missing data, risk prediction, machine learning, electronic health record data, random forests
39 40	Word count: 3,259
41 42	
43	
44 45	
46 47	
48	
49 50	
51	
52 53	
53 54	
55 56	
57	
58 59	1
60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### BMJ Open

#### ABSTRACT

2 3	<b>Objective:</b> There has been a proliferation of approaches to statistical methods and missing data imputation as electronic
4 5 6	health records become more plentiful; however, the relative performance on real-world problems is unclear.
7 8	Materials and Methods: Using 355,823 ICU hospitalizations at over 100 hospitals in the nationwide VA healthcare
9 10 11	system (2014-2017), we systematically varied 3 approaches: how we extracted and cleaned physiologic variables; how
12 13	we handled missing data (using mean value imputation, random forest, extremely randomized tress (extra-trees
14 15	regression), ridge regression, normal value imputation, and case-wise deletion); and how we computed risk (using
16 17 18	logistic regression, random forest, and neural networks). We applied these approaches in a 70% development sample
19 20	and tested the results in an independent 30% testing sample. Area under the ROC Curve (AUROC) was used to quantify
21 22	model discrimination.
23 24 25	Results: In 355,823 ICU stays, there were 34,867 deaths (9.8%) within 30 days of admission. The highest AUROC's
26 27	obtained for each primary classification method were very similar: 0.83 (95% CI [0.83-0.83]) to 0.85 (95% CI 0.840.85).
28 29	Likewise, there was relatively little variation within classification method by the missing value imputation method
30 31 32	used—except when case-wise deletion was applied for missing data.
33 34	<b>Conclusion:</b> Variation in discrimination was seen as a function of data cleanliness, with logistic regression suffering the
35 36 37	most loss of discrimination in the least clean data. Losses in discrimination were not present in random forest and neural
38 39	networks even in naively extracted data. Data from a large nationwide health system revealed interactions between
40 41 42	missing data imputation techniques, data cleanliness, and classification methods for predicting 30-day mortality.
42 43 44	
45 46	
47 48 49	
50 51	
52 53	
54 55 56	
57 58	2
59 60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### STRENGTHS AND LIMITATIONS OF THIS STUDY

- This study focuses on a large, real world dataset consisting of 355,823 ICU stays at over 100 different facilities. •
- Multiple methods of model fitting and missing data imputation were implemented in standardized ways that • reflect common practice.
- The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility, and we encourage validation on other datasets.
- Due to high dimensionality of method combinations, this study only considered one outcome, and only • considered one standardization method and decided upon an a priori approach within each dataset / categorization model / missingness imputation triad.

, retribution is used in the constraints of the con

#### **BMJ** Open

#### INTRODUCTION

Risk adjustment plays an increasingly central role in the organization, care of, and science about critically ill patients[1, 2]. Statistical adjustment, including the handling of missing data, is essential for many performance measurements as well as pay-for-performance and shared savings systems. It is used to stratify the care of patients for treatments and track quality improvement efforts over time[3]. It is routinely measured, even in clinical trials, to assess confounder balance between arms and may form part of RCT enrollment or drug approval criteria[4].

As a result, there has been a proliferation of risk scores and missing data imputation tools both for the common task of short-term mortality prediction and for more specialized tasks. Many statistical tools have been promoted. Rules of thumb have developed and existed long enough to be critiqued[5-9]. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines offer standardization of reporting[10]. Textbooks have emerged[11]. Yet questions remain on fundamental pragmatic issues: How clean does the data have to be to prevent the so-called "garbage in, garbage out (GIGO)" phenomenon? How sensitive are methods to missing data and how should it be handled? Do these analytic decisions interact?

To address such questions, we compared the performance of an array of methods on a single standardized problem—the prediction of 30-day mortality based on demographics, day 1 laboratory results, comorbidities, and diagnoses among patients admitted to the Intensive Care Unit (ICU) at any hospital in the nationwide Veterans Health Administration system[12-14]. Using the same set of real ICU admissions, we systematically varied three parameters: the approach used to extract and clean physiologic variables from the electronic health record; the approach used to handle missing data; and the approach used to compute the risk. We systematically applied these approaches in a 70% development sample and tested the results in an independent 30% testing sample, to provide real world comparisons to inform future pragmatic implementation of risk scores.

#### METHODS

#### Cohort

Data were drawn from the Veterans Affairs Patient Database (VAPD), which contains daily patient physiology for acute hospitalizations between January 1, 2014 and December 31, 2017. The VAPD includes patient demographics, laboratory results, and diagnoses that are commonly used to predict 30-day mortality from the day of admission. Here, we included data from all ICU hospitalizations on day 1 of each hospitalization. Full details of the VAPD have been published elsewhere[15].

The development of this database was reviewed and approved by the VA Ann Arbor Healthcare System's Institutional Review Board.

Four versions of the dataset were created for each hospitalization on admission: A) raw lab values extracted using only lab test names, B) raw lab values extracted using only Logical Observation Identifiers Names and Codes (LOINC), C) cleaned lab values extracted using both LOINC[16, 17] and searched text lab test names, and D) cleaned lab values converted to Acute Physiology And Chronic Health Evaluation (APACHE) points, extracted using both LOINC and lab test names.

#### No Patient and Public Involvement

This research was done without patient involvement. Patients were not invited to comment on the study design and were not consulted to develop patient relevant outcomes or interpret the results. Patients were not invited to contribute to the writing or editing of this document for readability or accuracy.

#### **Predictor Variables**

In our primary analyses, we adjust for 10 laboratory values that were collected within one day of hospital admission. Further patient-level adjustments included demographic characteristics (gender, age, race, and Hispanic ethnicity), 30 comorbidities, and 38 primary diagnoses. The individual comorbidities used in models are defined by methods described in van Walraven's implementation of the Elixhauser comorbidity score[18]. We adjust for 38 primary diagnoses drawn from the Healthcare Cost and Utilization (HCUP) Clinical Classification Software (CCS)[19], which consist of the top 20 most frequent single-level CCS diagnoses and 18 level-one multi-level categories of diagnoses (Appendix A.) In secondary

**BMJ** Open

analyses, to emphasize the role of data cleanliness, we estimate risk using *only* the laboratory values since the nonlaboratory values do not vary in data cleanliness and curation.

#### **Outcome Variable: 30-day mortality**

Our primary outcome variable is 30-day all-cause mortality, defined as death within 30 days of the admission date for the index hospitalization. Mortality is evaluated using the highly reliable Veterans Administration beneficiary death files which aggregate from multiple sources[12, 20, 21].

#### Statistical Analysis and Model Development

Random Forests is an ensemble machine learning method that aggregates the results of multiple decision trees fit on bootstrap samples of the original data[22, 23]. For each decision tree, the original data are bootstrapped to create a new dataset of the same size and the tree is fit to the new data. Instead of considering all predictors to determine the splitting criterion at a node, the split variable is chosen from a random subset of variables in order to reduce the correlation between different trees. Many such trees are grown, creating a 'forest'. Each observation is classified by each tree, and the majority classification over all trees is the predicted class. The ability of random forests to learn nonlinear and complex functions contributes to its predictive performance.

The neural network[24] can "learn" to classify samples without manual designed task-specific rules. The algorithm applies different weights to predictors and uses these transformations in subsequent "layers" of the neural net, culminating in the output layer with predictions. We applied the random forest and the neural network on our task. A traditional logistic regression model was also performed and compared.

Statistical analyses were performed with Python and the scikit-learn package[25].

Training and Testing Sets

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

The dataset was randomly split into a 70% training set and a 30% testing set. The same split was used for all classification methods. This process was replicated five times (five different training sets and corresponding testing set were generated),

and each time the models were fit on the training set and used to predict the 30-day mortality of the testing set.

#### **Missing Data and Imputation**

We imputed the missing values before training and testing the models, comparing:

- "Mean Value": the mean value of each variable in the training set was used to replace missing values[26].
- "Random Forest": used random forest to impute missing values (missForest)[27].
- "Extremely Randomized Trees (Extra-Trees Regression)": this method is similar to random forest but is faster[28, 29].
- "Ridge Regression": used Bayesian Ridge regression to impute missing values[30].
- "Normal Value"[31]: normal values were used to impute missing values—this is common in clinical prediction contexts in which it is assumed that clinicians order tests they fear are not normal, and therefore the absence of such a test is a sign that the clinician reviewed other aspects of the patient's case and judged the odds of physiologic abnormality so low that testing was not indicated.
- "No Missing": case-wise deletion[32].

#### Variable Importance and Partial Dependence Plots

Predictor variable importance was evaluated for random forests[33]. When classifying a sample using a decision tree, a predictor was used at each node. Predictors that appear more frequently and that reduce the misclassification more substantially are considered more important. By combining all trees in a random forest model, we assessed the variable importance of each predictor. Different values of the same predictor may have different effects on the prediction. We plotted the Partial Dependence Plots[30] to show how the value of predictors affects the prediction of 30-day mortality. Partial dependence plots were used to visualize non-linearity among variables.

The cohort comprised 355,823 ICU hospitalizations at over 100 different hospitals, as described elsewhere[15]. The mean

age of the cohort was 66.9 years, and there were 34,867 deaths within 30-days of admission, a primary outcome event

**ICU Only Cohort** 

355,823

66.9 (11.6)

341,579 (96.0)

256,293 (72.0)

73,855 (20.8)

25,675 (7.2)

20,532 (5.8)

34,867 (9.8)

9.5 (13.0)

#### RESULTS

**Cohort Description** 

rate of 9.8% (Table 1.)

Hospitalizations, N

Age, mean (SD), y

Black or African American

30-day Mortality, N (%)

Length of Stay, mean (SD), days

Variables

Male, N (%)

Race, N (%)

Hispanic, N (%)

White

Other

 Table 1. ICU Patient Demographics

#### 

Rates of data missingness for each laboratory value in each dataset are shown in Table 2. Dataset A has a high proportion of missing laboratory values for blood urea nitrogen (0.84) and hematocrit (0.85) compared to datasets B and C. This is due to dataset A using a single, broad lab test name to identify laboratory values: "BUN" for blood urea nitrogen and "hematocrit" for hematocrit. In contrast, datasets B and C incorporated LOINC codes for BUN and HCT, which result in

fewer missing laboratory values.

### Table 2. Proportion of Labs Missing

Using all Data for Model Development

Dataset	Albumin (albval)	Bilirubin (bili)	Blood urea nitrogen (bun)	a Creatnine (creat)	Glucose (glucose)	Hematocrit (hct)	Partial Presssure (pao2)	pH (pa)	Sodium (na)	White Blood Cell (wbc)
A	0.39	0.42	0.84	0.13	0.07	0.85	0.66	0.14	0.11	0.13
в	0.38	0.42	0.13	0.13	0.06	0.12	0.65	0.44	0.11	0.13
с	0.39	0.45	0.13	0.12	0.06	0.11	0.69	0.64	0.11	0.13

Figure 1 shows the AUC scores of different classification models and imputation methods in the primary analysis. The highest AUC's obtained for each primary classification method (rows of the figure: logistic regression, random forest, or a neural network) were very similar: AUC's of 0.83 to 0.85. Likewise, there was relatively little variation within classification method by the missing value imputation method used, be it mean value imputation, random forest, extremely randomized trees (extra-trees regression), ridge regression, or normal value imputation. All models suffered dramatic losses in discrimination when case-wise deletion was used for missing data in the least clean dataset (far right columns). Full model performance for each condition can be seen in Appendix B.

Variation in discrimination was seen, however, across classification methods, as a function of data cleanliness. (Note that the analyst was blinded during the analysis to how each dataset was developed, and hence did not know which was "cleanest"). In the logistic regression model developed using the least clean data (dataset A had raw lab values extracted using only lab test names), performance was always lower than the performance with the more complete and clean datasets—by AUC's of 0.05 to about 0.1, p-value < 0.05). Similarly, performance in dataset B (extracted using LOINC codes without unit standardization) was lower and more unstable for mean value imputation and ridge regression. In marked contrast, neither random forests nor neural networks showed such reduced performance when developed in less clean data—in no case did the AUC degradations exceed 0.025 despite similar optimal performance.

#### Secondary Analysis Using only Laboratory Values

The primary analysis presented above considers the real world case in which demographics, diagnoses, and laboratory values are used in combination with risk model prediction. Yet, of these, only laboratory values were subject to variation in cleanliness. We, therefore, conducted a secondary analysis using only laboratory values to assess more clearly the impact of data quality. Results are shown in Figure 2.

Average model performance with this much smaller group of predictors is, as expected, somewhat lower with less data optimal AUC's typically range from 0.73 to 0.78 across combinations of classification model and missing data imputation. No uniformly superior strategy is evident, save markedly lower performance of case-wise deletion in the least clean

#### **BMJ** Open

dataset (A). As before, logistic regression shows markedly reduced discrimination when developed in the least clean data set. Neural networks show consistent performance.

Also notable is the marked reduction of discrimination of random forest models and neural network models regardless of the missing data imputation model used within dataset D. Dataset D has the "cleanest" data, in that it has hand-curated inclusion criteria, standardization of units, and conversion of values from their continuous scale to a semi-quantitative set of "points" as is done in the APACHE scoring algorithms. Attempting to work with such standardized point values as inputs consistently resulted in markedly worse discrimination in random forest models and neural network models than using other "less clean" datasets (the difference between Dataset D and other datasets is significant with a p-value < 0.05).

#### Variable Importance

The most important predictors of 30-day mortality were age and laboratory values. Age had the highest importance scores, regardless of which dataset was used, indicating that age is the most important variable when predicting 30-day mortality. The 10 laboratory values also had high importance scores. For datasets A, B, and C, laboratory values fell in the top-13 most important variables, and there were at least 8 laboratory values in the top-10 most important variables. However, for dataset D, there were only 6 laboratory values in the top-10 most important variables, and the variable white blood cell score ranked 20<sup>th</sup>. This may indicate that transforming laboratory values to APACHE scores results in the loss of information contained in the original values and negatively influences the performance of the random forest model.

#### **Partial Dependence Plots**

As it is hard to visualize the relationship between multiple predictors and the outcome, we created partial dependence plots to show the effect of predictors on the outcome [34]. The plots can also show whether the relationship between a specific predictor and the outcome is linear, guadratic, monotonic, or more complex. Further analysis can be done by combining the partial dependence plots and medical knowledge. Figure 3 and Figure 4 are the partial dependence plots for the pH score and the  $PaO_2$  score. We will take these as examples to show how the value of predictors in different datasets affects 30-day mortality. The X-axis is the value of the predictor. For each value of the predictor, the Y-axis is the averaged model output for all observations with the corresponding value of the predictor. A higher partial dependence

value corresponds to a higher risk of mortality. As we know, the normal value of the pH score is 7.4, and both higher values and lower values are abnormal. Typically, abnormal values lead to a larger risk of death. Therefore, a U-shaped partial dependence plot is to be expected for datasets A, B, and C. However, only the plot for dataset C is U-shaped. This is because dataset C is "cleaner" than datasets A and B, and the models can learn the real effect of pH score on 30-day mortality. Datasets A and B are not as clean as dataset C, as some other variables are presented in these datasets as pH score. Thus, it is difficult for the models to utilize the pH score variable in datasets A and B. This result indicates that cleaner variables benefits the classification models. However, not all variables have this problem. For most other variables such as the PaO<sub>2</sub> score, the plots of datasets A, B, and C have similar trends.

#### DISCUSSION

We used real data from a large nationwide health system to explore the interaction between missing data imputation techniques, data cleanliness, and classification methods for the common problem of predicting 30-day mortality in a holdout testing dataset. In brief, we found that any of several imputation techniques other than case-wise deletion performed equivalently in terms of discrimination, regardless of data cleanliness or classification method used. We found that logistic regression showed worse discrimination with less carefully cleaned data than did random forest or neural networks. Random forest models (and to a degree, neural networks) displayed diminished discrimination when given data that had been too highly cleaned and standardized prior to use.

#### Relationship to Past Research

Missing data are ubiquitous in large datasets. Even when missingness is completely at random, missing data lead to significant loss in statistical power and predictive ability[32]. We have previously found that the Random Forest method consistently produced the lowest imputation error compared to commonly used imputation methods[26]. Random Forest had the smallest prediction difference when 10-30% of the laboratory data was missing. Our present analysis of real data shows that as more specialized laboratory values are introduced into the prediction setting, much higher levels of missingness may be present. We thereby extend the previous finding that Random Forest continues to perform well for

#### **BMJ** Open

missing data. Our findings on the poor performance of case-wise deletion as an approach to handling missing data are in agreement with mainstream recommendations for more than two decades[32].

Our findings on missing data are of note because of the distinctive, yet real world, way in which missing data were generated. There were two missingness processes. First, clinicians in routine practice only sometimes order any given laboratory, and thus the presence or absence of an order may itself provide prognostic importance. [35] Second, an effort to identify all target laboratory values may or may not succeed. Even in a large system with a strong tradition of centralization, laboratory labeling practices vary over time and clinical insight is often necessary to distinguish valid laboratory tests[36]. For any given data pull, it is not trivial to understand which missing values represent failure to find data that exist versus representing true missingness. Past work has rarely explicitly considered these distinct missingness-generating processes (in addition to true missingness at random) at their distinct implications.

The finding of poorer discrimination of Random Forest in models where the data were fully standardized and cleaned was not anticipated given past literature. The APACHE score was designed to simplify the lab results and to help doctors predict mortality [2]. Even in its more recent incarnations, APACHE transforms continuous lab results into discrete acute physiology scores[37]. Our data suggest that transforming lab results to APACHE scores is not necessary for Random Forest and may even lead to the loss of information[23]. Remarkably, even standardization to equivalent units across institutions may not be necessary—but at the same time, this means that sources of variance other than simply the laboratory value may also be subtly incorporated into risk-prediction with non-standardized ways. It is a case-specific decision as to whether incorporation of such variance is helpful for a given task or is a source of bias.

#### Implications

Our findings have implications for both practitioners seeking to implement a given prediction rule and scientists interested in risk-prediction generally. For practitioners, no given method yields consistently superior results in terms of discrimination. Therefore, other performance considerations, whether psychometric or implementation ease, may play

an important role. They also suggest that missing data imputation approaches other than case-wise deletion during development are mandatory.

Our results also note that Random Forests and neural networks were strikingly robust to even quite naively prepared data, in contrast to logistic regression. This suggests that the truth of the oft-quoted aphorisms about "garbage in, garbage out" may depend on the categorization model and missing data imputation method used. In situations where ascertainment and cleaning of data are more costly, random forests may offer pragmatic advantages if these findings are replicable.

#### Strengths and Limitations

Strengths of our analysis include its use of real world data, with real world data generation and missingness-generation problems on an established problem encountered by medical researchers and clinicians. We also used multiple methods implemented in standardized ways. The approach we used for each implementation is available in an Appendix or via GitHub to allow transparency and reproducibility.

Limitations of our analysis stem fundamentally from the nearly infinite combinations of analysis factors that might be varied, and our inability to explore such a high dimensional space. Thus we only considered one outcome and one standardization method, and decided upon an a priori approach for each combination of dataset, categorization model, and missingness imputation method used. Other outcomes and other possible data structures (such as using trends in data) may yield different answers. We focus on discrimination, as measured by AUC, but other measurement properties are assuredly also important. We also focused on individual-level prediction, as opposed to considering the impact on hospital-level quality assessment or other tasks for which these results may be used.

#### CONCLUSION

In sum, our results suggest that there is little variation in discrimination among different statistical classification models in well-cleaned data using modern missing data imputation techniques. As such, the decision about which of the well-

#### Page 15 of 34

#### BMJ Open

1	performing imputation and adjustment methods to use can be made based on other factors relevant to the particular
2 3	application—as long as the lower performing methods are avoided. If these findings are replicated in other data with other
54 55	
56 57	
58 59	14
60	For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

#### **Figure Captions**

- Figure 1. AUC Scores, Full Model
- Figure 2. AUC Scores for lab-only predictors
- Figure 3. Partial Dependence Plots for pH
- Figure 4. Partial Dependence Plots for PaO2

torbeer terien only

#### ACKNOWLEDGEMENTS

#### Author Contributions

**Theodore J. Iwashyna:** Conceptualization, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing

Cheng Ma: Formal analysis, software, visualization, writing-original draft, and writing-review & editing

Xiao Qing Wang: Data curation; Writing – original draft; Writing – review & editing

Sarah Seelye : Writing – original draft; Writing – review & editing

Ji Zhu : Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

Akbar K. Waljee: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing

**Declarations of funding interests:** This work was supported by a career development grant award [CDA 11-217 to A.K.W] and a Merit Review Award Number [IIR 16-024 to A.K.W, 17-045 to T.J.I.] from the United States Department of Veterans Affairs Health Services Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Michigan, the Veterans Affairs, the U.S. Government, or the National Institutes of Health.

Disclosures: The authors declare there are no conflicts to disclose

**Guarantor of the article:** Akbar K. Waljee the submission's guarantor, takes responsibility for the integrity of the work as a whole, from inception to published article.

#### Licensing Statement:

I, Akbar Waljee, as the Submitting Author have the right to grant and does grant on behalf of all authors of the Work (as defined in the below author license), an exclusive license and/or a non-exclusive license for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY license shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in BMJ Open and any other BMJ products and to exploit all rights, as set out in our <u>license</u>.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons license – details of these licenses and which <u>Creative Commons</u> license will apply to this Work are set out in our license referred to above.

## 53 Data Sharing Statement:54

Appendices and statistical code are available via Github at <a href="https://github.com/CCMRcodes/GIVO">https://github.com/CCMRcodes/GIVO</a>. The dataset cannot be
 disseminated due to inclusion of sensitive patient information under VA regulations.

 tor peer terier only

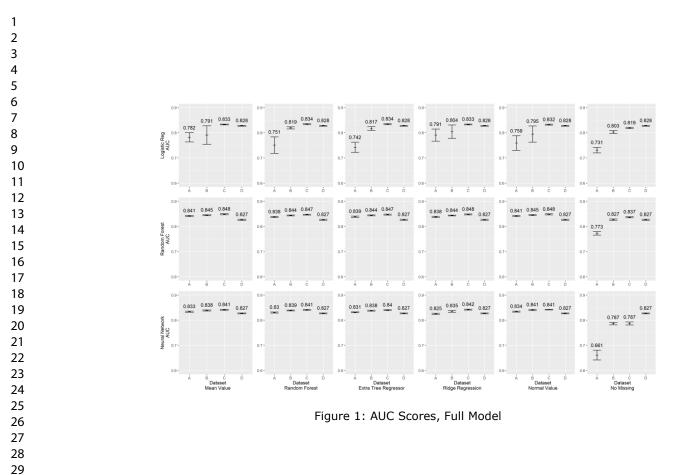
#### REFERENCES

- [1] Iezzoni LI. Risk adjustment for measuring health care outcomes. 4th ed. Chicago, Ill. Arlington, VA: Health
   Administration Press; AUPHA; 2013.
- [2] Lane-Fall MB, Neuman MD. Outcomes measures and risk adjustment. Int Anesthesiol Clin. 2013;51:10-21.
- [2] Lane-Fairwis, Neuman WD. Outcomes measures and risk adjustment. Int Anestnesion (
   [3] Quality AfHRa. Part II. Introduction to Measures of Quality (continued). Rockville, MD.
- [4] Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A, et al. Efficacy and safety of
- 8 recombinant human activated protein C for severe sepsis. N Engl J Med. 2001;344:699-709.
- [5] Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction.
  Stat Med. 1984;3:143-52.
- [6] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions
   and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
- [7] van Smeden M, de Groot JA, Moons KG, Collins GS, Altman DG, Eijkemans MJ, et al. No rationale for 1 variable per 10
   events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163.
- [8] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a
- <sup>17</sup> multivariable prediction model: PART II binary and time-to-event outcomes. Stat Med. 2019;38:1276-96.
- [9] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic
   prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2019;28:2455-74.
- [10] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a
- multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern
   Med. 2015;162:W1-73.
   Med. 2015;162:W1-73.
- [11] Steyerberg EW, ProQuest (Firm). Clinical prediction models a practical approach to development, validation, and
   updating. New York: Springer; 2009. p. xxviii, 497 p.
- [12] Render ML, Kim HM, Welsh DE, Timmons S, Johnston J, Hui S, et al. Automated intensive care unit risk adjustment: results from a National Veterans Affairs study. Crit Care Med. 2003;31:1638-46.
- [13] Render ML, Deddens J, Freyberg R, Almenoff P, Connors AF, Wagner D, et al. Veterans Affairs intensive care unit risk
   adjustment model: validation, updating, recalibration. Crit Care Med. 2008;36:1031-42.
- <sup>30</sup> [14] Render ML, Freyberg RW, Hasselbeck R, Hofer TP, Sales AE, Deddens J, et al. Infrastructure for quality
- transformation: measurement and reporting in veterans administration intensive care units. BMJ Qual Saf. 2011;20:498-507.
- [15] Wang XQ, Vincent BM, Wiitala WL, Luginbill KA, Viglianti EM, Prescott HC, et al. Veterans Affairs patient database (VAPD 2014-2017): building nationwide granular data for clinical discovery. DMC Model Dev Market Dev Market Dev
- (VAPD 2014-2017): building nationwide granular data for clinical discovery. BMC Med Res Methodol. 2019;19:94.
- [16] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying
   laboratory observations: a 5-year update. Clin Chem. 2003;49:624-33.
- 38 [17] Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and
- codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results.
   Clin Chem. 1996;42:81-90.
- [18] van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures
   inter a neint system for hearing a derivative state. Mad. Comp. 2000; 17:222-22
- into a point system for hospital death using administrative data. Med Care. 2009;47:626-33.
   [40] UCUP US Tools & Software Page 2010.
- [19] HCUP-US Tools & Software Page. 2019.
- [20] Hooper TI, Gackstetter GD, Leardmann CA, Boyko EJ, Pearse LA, Smith B, et al. Early mortality experience in a large
   military cohort and a comparison of mortality data sources. Popul Health Metr. 2010;8:15.
- [21] Prescott HC, Kepreos KM, Wiitala WL, Iwashyna TJ. Temporal Changes in the Influence of Hospitals and Regional
- 48 Healthcare Networks on Severe Sepsis Mortality. Crit Care Med. 2015;43:1368-74.
- 49 [22] Breiman L. Classification and regression trees. New York, NY: Chapman & Hall; 1993.
- [23] Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
   [24] Omiduar O. Dayhoff JE. Science Direct (Online convice). Neural
- [24] Omidvar O, Dayhoff JE, ScienceDirect (Online service). Neural networks and pattern recognition. San Diego, Calif.:
   Academic Press; 1998.
   [25] Dedregees E. Veregeugung C. Complete A. Michael V. Thinkson D. Calif. I. Status and San Diego. Calif.
- [25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python.
   Journal of machine learning research. 2011;12:2825–30.
- 56
- 57
- 58 59

[26] Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3.

- [27] Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data.
- Bioinformatics. 2012;28:112-8.

- [28] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63:3-42.
- [29] Maree R, Geurts P, Wehenkel L. Random subwindows and extremely randomized trees for image classification in cell biology. BMC Cell Biol. 2007;8 Suppl 1:S2.
- [30] Hastie T, Friedman J, Tibshirani R, SpringerLink (Online service). The Elements of Statistical Learning Data Mining, Inference, and Prediction. New York, NY: Springer New York : Imprint: Springer; 2001.
- [31] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health
- evaluation: a physiologically based classification system. Crit Care Med. 1981;9:591-7.
- [32] Allison PD, Sage Publications. Missing data. Thousand Oaks, [Calif.] ; London: SAGE; 2002. p. 1 online resource (vi, 91 p.).
- [33] Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees.
- Proceedings of the 26th International Conference on Neural Information Processing Systems Volume 1. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. p. 431-9.
- [34] Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics. 2001;29:1189-232.
- [35] Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.
- [36] Wiitala WL, Vincent BM, Burns JA, Prescott HC, Waljee AK, Cohen GR, et al. Variation in Laboratory Test Naming
- Conventions in EHRs Within and Between Hospitals: A Nationwide Longitudinal Study. Med Care. 2019;57:e22-e7.
- [37] Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV:
- hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34:1297-310.



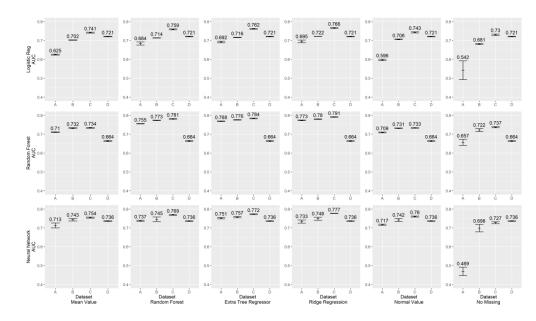
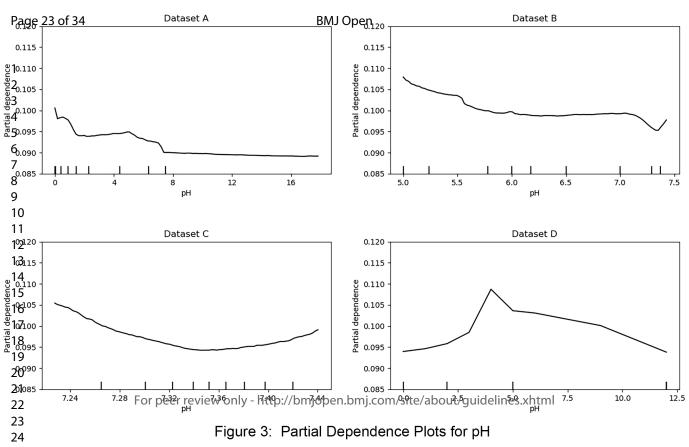
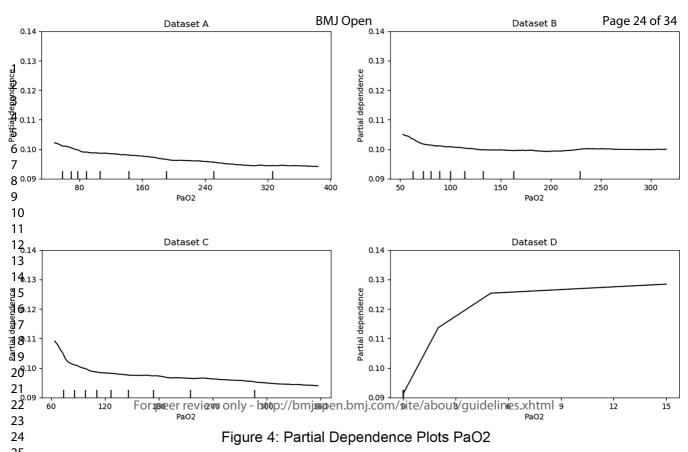


Figure 2. AUC Scores for lab-only predictors





#### Appendix A. Patient-level variables included in models

Demographics	Gender, Age, Race (White, Black or African American, Asian, Native
	Hawaiian or other Pacific Islander, Unknown), Hispanic ethnicity
	Hypertension, Congestive Heart Failure, Cardiac Arrhythmia, Valvular
	Disease, Pulmonary Circulation Disorders, Peripheral Vascular
	Disorders, Paralysis, Other Neurological Disorders, Chronic Pulmonary
Comorbidities,	Disease, Diabetes Uncomplicated, Diabetes Complicated,
included in Elixhauser	Hypothyroidism, Renal Failure, Liver Disease, Peptic Ulcer Disease
	excluding bleeding, AIDS/HIV, Lymphoma, Metastatic Cancer, Solid
	Tumor without Metastasis, Rheumatoid Arthritis/Collagen,
	Coagulopathy, Obesity, Weight Loss, Fluid and Electrolyte Disorders,
	Blood Loss Anemia, Deficiency Anemia, Alcohol Abuse, Drug Abuse,
	Psychoses, Depression
	Top 20 most frequent single-level CCS diagnoses:
	Congestive Heart Failure (non-hypertensive), Non-specific Chest Pain,
	Coronary Atherosclerosis and Other Heart Disease, Cardiac
	Dysrhythmias, Alcohol-related Disorders, Septicemia (except in labor),
	Chronic Obstructive Pulmonary Disease and Bronchiectasis,
	Pneumonia, Skin and Subcutaneous Tissue Infections, Osteoarthritis,
	Complication of Device (implant or graft), Complications of Surgical
	Procedures or Medical Care, Diabetes Mellitus with Complications,
	Respiratory Failure, Urinary Tract Infections, Renal Failure,
Diagnoses,	Spondylosis, Acute Myocardial Infarction, Fluid and Electrolyte
HCUP CCS single-level and multi-level	Disorders, Gastrointestinal Hemorrhage
	bisorders, edistrointestinar nemorninge
	18 level 1 multi-level CCS categories:
	Infectious and Parasitic Diseases, Neoplasms, Endocrine Disorders,
	Anemia, Mental Illness, Diseases of the Nervous System, Diseases of
	the Circulatory System, Diseases of the Respiratory System, Diseases of
	the Digestive System, Diseases of the Genitourinary System,
	Complications of Pregnancy or Childbirth, Skin Disease, Diseases of the
	Musculoskeletal System, Congenital Anomalies, Perinatal Conditions, Injury and Poisoning, Other Health Status Conditions, Other Residual
	Codes
	Albumin, Bilirubin, Blood Urea Nitrogen, Creatinine, Glucose,
	Abamin, binabin, bioda orea Micogen, creatinine, diacose,
Laboratory values	Hematocrit, Partial pressure of oxygen score, pH score, Sodium, White

#### Appendix B

#### Table B.1: Model Performances (Full Model)

Classification		Imputation	AUROC	Optimal	Predicted Cases		es Accurate Rate				Brier
Method	Dataset	Method	(95%CI)	Cutoff	Death	Survival	Death	Survival	Sensitivity	Specificity	Score
Logistic Regression	A	Mean Value	0.78(0.76- 0.80)	0.48	37199	69549	0.21	0.96	0.74	0.69	0.19
Logistic Regression	В	Mean Value	0.79(0.75- 0.83)	0.46	34915	71833	0.24	0.97	0.80	0.72	0.17
Logistic Regression	С	Mean Value	0.83(0.83- 0.83)	0.46	35970	70778	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Random Forest	0.75(0.72- 0.78)	0.49	39528	67220	0.18	0.95	0.69	0.66	0.21
Logistic Regression	В	Random Forest	0.82(0.82- 0.82)	0.49	31996	74752	0.25	0.97	0.77	0.75	0.17
Logistic Regression	С	Random Forest	0.83(0.83- 0.84)	0.46	36017	70731	0.23	0.97	0.79	0.71	0.17
Logistic Regression	A	Extra Trees Regression	0.74(0.72- 0.76)	0.46	45762	60986	0.17	0.96	0.74	0.61	0.21
Logistic Regression	В	Extra Trees Regression	0.82(0.81- 0.82)	0.47	33642	73106	0.25	0.97	0.79	0.74	0.17
Logistic Regression	С	Extra Trees Regression	0.83(0.83- 0.84)	0.47	34579	72169	0.24	0.97	0.78	0.73	0.17
Logistic Regression	A	Ridge Regression	0.79(0.77- 0.82)	0.46	38579	68169	0.21	0.97	0.78	0.68	0.18
Logistic Regression	В	Ridge Regression	0.80(0.78- 0.83)	0.46	35034	71714	0.24	0.97	0.80	0.72	0.17

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	27	of	34
------	----	----	----

		<b>D</b> : 1	0.00/0.00								
Logistic Regression	С	Ridge Regression	0.83(0.83- 0.84)	0.47	35220	71528	0.23	0.97	0.78	0.72	0.17
Logistic Regression	A	Normal Value	0.76(0.73- 0.79)	0.50	37392	69356	0.18	0.95	0.63	0.68	0.22
Logistic Regression	В	Normal Value	0.80(0.76- 0.83)	0.50	30977	75771	0.26	0.97	0.76	0.76	0.17
Logistic Regression	С	Normal Value	0.83(0.83- 0.83)	0.46	35676	71072	0.23	0.97	0.78	0.72	0.17
Logistic Regression	A	No Missing	0.73(0.72- 0.74)	0.43	37658	69090	0.18	0.95	0.66	0.68	0.18
Logistic Regression	В	No Missing	0.8(0.80- 0.81)	0.42	33153	73595	0.24	0.97	0.76	0.74	0.14
Logistic Regression	С	No Missing	0.82(0.82- 0.82)	0.44	35333	71415	0.22	0.96	0.76	0.72	0.16
Logistic Regression	D	None	0.83(0.83- 0.83)	0.47	34184	72564	0.24	0.97	0.78	0.73	0.17
Random Forest	A	Mean Value	0.84(0.84- 0.84)	0.12	32330	74418	0.26	0.97	0.79	0.75	0.07
Random Forest	В	Mean Value	0.85(0.84- 0.85)	0.11	32642	74106	0.26	0.97	0.80	0.75	0.07
Random Forest	С	Mean Value	0.85(0.85- 0.85)	0.11	33548	73200	0.25	0.97	0.81	0.74	0.07
Random Forest	A	Random Forest	0.84(0.84- 0.84)	0.12	32659	74089	0.25	0.97	0.78	0.75	0.07
Random Forest	В	Random Forest	0.84(0.84- 0.85)	0.11	34093	72655	0.25	0.97	0.81	0.73	0.07

**BMJ** Open

Random		Random	0.85(0.85-	0.11	33029	73719	0.25	0.97	0.80	0.74	0.0
Forest	С	Forest	0.85)	0.11	55029	/3/19	0.25	0.97	0.80	0.74	0.0
Random		Extra Trees	0.84(0.84-	0.11	32938	73810	0.25	0.97	0.79	0.74	0.0
Forest	А	Regression	0.84)	0.11	32938	/3810	0.25	0.97	0.79	0.74	0.0
Random		Extra Trees	0.84(0.84-	0.12	22411	74227	0.26	0.07	0.80	0.75	0.0
Forest	В	Regression	0.85)	0.12	32411	74337	0.26	0.97	0.80	0.75	0.0
Random		Extra Trees	0.85(0.85-	0.11	33567	73181	0.25	0.97	0.80	0.74	0.0
Forest	С	Regression	0.85)	0.11	55507	/5101	0.25	0.97	0.80	0.74	0.0
Random		Ridge	0.84(0.45-	0.11	34587	72161	0.24	0.97	0.80	0.73	0.0
Forest	A	Regression	0.84)		54507	/2101	0.24	0.57	0.00	0.75	0.0
Random		Ridge	0.84(0.84-	0.12	31643	75105	0.26	0.97	0.79	0.76	0.0
Forest	В	Regression	0.85)	0.12	51045	/ 5105	0.20	0.57	0.75	0.70	0.0
Random		Ridge	0.85(0.85-	0.12	32531	74217	0.25	0.97	0.79	0.75	0.0
Forest	С	Regression	0.85)	0.12	52551		0.23	0.57	0.75	0.75	0.0
Random		Normal	0.84(0.84-	0.12	31234	75514	0.26	0.97	0.78	0.76	0.0
Forest	A	Value	0.84)	0.12	51254	75514	0.20	0.57	0.70	0.70	0.0
Random		Normal	0.85(0.84-	0.11	32711	74037	0.26	0.97	0.80	0.75	0.0
Forest	В	Value	0.85)	0.11	52711	74037	0.20	0.57	0.00	0.75	0.0
Random		Normal	0.85(0.85-	0.12	31159	75589	0.26	0.97	0.78	0.76	0.0
Forest	С	Value	0.85)	0.12	51155	75505	0.20	0.57	0.70	0.70	0.0
Random			0.77(0.77-	0.18	36332	70416	0.20	0.96	0.71	0.70	0.0
Forest	А	No Missing	0.78)	0.10	50552	/0410	0.20	0.90	0.71	0.70	0.0
Random			0.83(0.82-	0.16	31836	74912	0.25	0.97	0.77	0.75	0.0
Forest	В	No Missing	0.83)	0.10	21020	/4J12	0.25	0.57	0.77	0.75	0.0
Random			0.84(0.83-	0.16	34517	72231	0.24	0.97	0.78	0.73	0.0
Forest	С	No Missing	0.84)	0.10	5451/	12231	0.24	0.97	0.76	0.75	0.0

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 29 of 34

B	MJ Open	

Random			0.83(0.83-	0.11	33407	73341	0.24	0.97	0.77	0.74	
Forest	D	None	0.83)	0.11	00107	/0011	0.21	0.07	0177	0.71	
Neural		Mean	0.83(0.83-	0.53	35031	71717	0.24	0.97	0.80	0.72	
Network	А	Value	0.84)	0.55	55051	, 1, 1,	0.24	0.57	0.00	0.72	
Neural		Mean	0.84(0.84-	0.52	32716	74032	0.25	0.97	0.79	0.75	
Network	В	Value	0.84)	0.52	52710	74032	0.25	0.57	0.75	0.75	
Neural		Mean	0.84(0.84-	0.53	32549	74199	0.25	0.97	0.79	0.75	
Network	С	Value	0.84)	0.55	52549	74199	0.25	0.97	0.79	0.75	
Neural		Random	0.83(0.83-	0.55	32515	74233	0.25	0.97	0.77	0.75	
Network	А	Forest	0.83)	0.55	52515	74255	0.25	0.97	0.77	0.75	
Neural		Random	0.84(0.84-	0.57	30842	75906	0.26	0.97	0.77	0.76	
Network	В	Forest	0.84)	0.57	50042	75500	0.20	0.57	0.77	0.70	
Neural		Random	0.84(0.84-	0.55	34144	72604	0.24	0.97	0.79	0.73	
Network	С	Forest	0.84)	0.55	54144	72004	0.24	0.57	0.75	0.75	
Neural		Extra Trees	0.83(0.83-	0.50	37351	69397	0.23	0.97	0.82	0.70	
Network	А	Regression	0.83)	0.50	57551	05557	0.25	0.57	0.02	0.70	
Neural		Extra Trees	0.84(0.84-	0.54	33529	73219	0.25	0.97	0.80	0.74	
Network	В	Regression	0.84)	0.54	33329	73219	0.25	0.97	0.80	0.74	
Neural		Extra Trees	0.84(0.84-	0.49	33324	73424	0.25	0.97	0.79	0.74	
Network	С	Regression	0.84)	0.45	55524	73424	0.25	0.57	0.75	0.74	
Neural		Ridge	0.83(0.82-	0.51	33864	72884	0.24	0.97	0.78	0.73	
Network	А	Regression	0.83)	0.51	55004	/2004	0.24	0.97	0.70	0.75	
Neural		Ridge	0.84(0.83-	0.52	31186	75562	0.26	0.97	0.78	0.76	(
Network	В	Regression	0.84)	0.52	21190	10002	0.20	0.97	0.78	0.70	
Neural		Ridge	0.84(0.84-	0.55	22145	74602	0.25	0.07	0.79	0.75	
Network	С	Regression	0.84)	0.55	32145	74603	0.25	0.97	0.78	0.75	

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	30 of	34
------	-------	----

									0		
Network	D	None	0.83)			Q	10				
Neural		Nono	0.83(0.83-	0.52	33760	72988	0.24	0.97	0.78	0.73	0.18
Neural Network	С	No Missing	0.79(0.78- 0.79)	0.59	38619	68129	0.20	0.96	0.75	0.68	0.24
Neural Network	В	No Missing	0.79(0.78- 0.79)	0.59	32676	74072	0.23	0.96	0.71	0.74	0.21
Neural Network	A	No Missing	0.66(0.64- 0.68)	0.76	49011	57737	0.14	0.94	0.65	0.56	0.50
Neural Network	С	Normal Value	0.84(0.84- 0.84)	0.55	32864	73884	0.25	0.97	0.79	0.75	0.18
Neural Network	В	Normal Value	0.84(0.84- 0.84)	0.45	35026	71722	0.24	0.97	0.81	0.72	0.16
Neural Network	А	Normal Value	0.83(0.83- 0.84)	0.58	31675	75073	0.25	0.97	0.77	0.76	0.19

Classification		Imputation	AUROC	Optimal	Predict	ed Cases	Accura	ite Rate			Brie
Method	Dataset	Method	(95%CI)	Cutoff	Death	Survival	Death	Survival	Sensitivity	Specificity	Sco
Logistic Regression	A	Mean Value	0.63(0.62- 0.63)	0.50	32327	74421	0.16	0.93	0.50	0.72	0.24
Logistic Regression	В	Mean Value	0.70(0.70- 0.70)	0.47	33350	73398	0.21	0.95	0.66	0.73	0.20
Logistic Regression	С	Mean Value	0.74(0.74- 0.74)	0.47	35248	71500	0.19	0.95	0.62	0.70	0.22
Logistic Regression	A	Random Forest	0.68(0.68- 0.69)	0.48	39566	67182	0.17	0.94	0.63	0.66	0.23
Logistic Regression	В	Random Forest	0.71(0.71- 0.71)	0.45	37758	68990	0.20	0.96	0.71	0.69	0.20
Logistic Regression	С	Random Forest	0.76(0.76- 0.76)	0.47	38421	68327	0.18	0.95	0.66	0.67	0.21
Logistic Regression	A	Extra Trees Regression	0.69(0.69- 0.70)	0.46	43607	63141	0.17	0.95	0.69	0.62	0.22
Logistic Regression	В	Extra Trees Regression	0.72(0.72- 0.72)	0.44	39295	67453	0.20	0.96	0.73	0.67	0.19
Logistic Regression	С	Extra Trees Regression	0.76(0.76- 0.76)	0.45	42675	64073	0.17	0.95	0.71	0.63	0.21
Logistic Regression	A	Ridge Regression	0.70(0.69- 0.70)	0.47	42514	64234	0.17	0.95	0.69	0.63	0.22
Logistic Regression	В	Ridge Regression	0.72(0.72- 0.72)	0.44	39856	66892	0.20	0.96	0.75	0.67	0.19

Page	32	of	34
· ~	-	•••	•

Logistic		Ridge	0.77(0.76-	0.45	42737	64011	0.17	0.95	0.71	0.63	0.2
Regression	С	Regression	0.77)								
Logistic		Normal	0.60(0.59-	0.49	31990	74758	0.15	0.92	0.46	0.72	0.2
Regression	А	Value	0.60)	0.45	51550	/4/30	0.15	0.52	0.40	0.72	0.2
Logistic		Normal	0.71(0.70-	0.45	38325	68423	0.19	0.95	0.70	0.68	0.2
Regression	В	Value	0.71)	0.45	30323	00425	0.19	0.95	0.70	0.08	0.2
Logistic		Normal	0.74 (0.74-	0.46	41447	65301	0.17	0.95	0.68	0.64	0.2
Regression	С	Value	0.75)	0.40	41447	05301	0.17	0.95	0.08	0.64	0.2
Logistic			0.54 (0.49-	0.57	17678	89070	0.15	0.91	0.25	0.84	0.2
Regression	A	No Missing	0.59)		1,0,0	00070	0.15	0.51	0.20	0.04	0.2
Logistic			0.68 n(0.68-	0.45	32766	73982	0.20	0.95	0.64	0.73	0.1
Regression	В	No Missing	0.68)	0.45	32700	/ 5502	0.20	0.55	0.04	0.75	0.1
Logistic			0.73(0.73-	0.50	30965	75783	0.19	0.94	0.55	0.74	0.2
Regression	С	No Missing	0.73)	0.50	30303	/3/03	0.13	0.54	0.55	0.74	0.2
Logistic			0.72(0.72-	0.49	35766	70982	0.19	0.95	0.64	0.70	0.2
Regression	D	None	0.72)	0.45	55700	70502	0.15	0.55	0.04	0.70	0.2
Random		Mean	0.71(0.71-	0.09	46226	60522	0.17	0.95	0.73	0.60	0.0
Forest	А	Value	0.71)	0.09	40220	00322	0.17	0.95	0.75	0.00	0.0
Random		Mean	0.73(0.73-	0.10	44628	62120	0.18	0.96	0.75	0.62	0.0
Forest	В	Value	0.73)	0.10	44020	02120	0.10	0.50	0.75	0.02	0.0
Random		Mean	0.73(0.73-	0.10	44525	62223	0.18	0.96	0.75	0.62	0.0
Forest	С	Value	0.74)	0.10	44323	02223	0.10	0.90	0.75	0.02	0.0
Random		Random	0.76(0.75-	0.09	41715	65033	0.18	0.96	0.74	0.65	0.0
Forest	А	Forest	0.76)	0.09	41/13	05033	0.18	0.90	0.74	0.05	0.0
Random		Random	0.77(0.77-	0.10	38154	68594	0.20	0.96	0.75	0.69	0.0
Forest	В	Forest	0.77)	0.10	50154	00334	0.20	0.90	0.75	0.05	0.0

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page 33	of 34
---------	-------

Random Forest	С	Random Forest	0.78(0.78- 0.78)	0.09	40709	66039	0.19	0.96	0.75	0.66	0.08
Random Forest	А	Extra Trees Regression	0.77(0.77- 0.77)	0.09	43230	63518	0.19	0.96	0.77	0.64	0.08
Random Forest	В	Extra Trees Regression	0.78(0.77- 0.78)	0.10	38734	68014	0.20	0.96	0.76	0.68	0.08
Random Forest	С	Extra Trees Regression	0.78(0.78- 0.79)	0.10	38810	67938	0.20	0.96	0.74	0.68	0.08
Random Forest	А	Ridge Regression	0.77(0.77- 0.78)	0.10	39913	66835	0.20	0.96	0.75	0.67	0.08
Random Forest	В	Ridge Regression	0.78(0.78- 0.78)	0.09	39663	67085	0.20	0.97	0.77	0.67	0.08
Random Forest	С	Ridge Regression	0.79(0.79- 0.79)	0.09	40249	66499	0.20	0.96	0.76	0.66	0.08
Random Forest	А	Normal Value	0.71(0.71- 0.71)	0.10	46047	60701	0.17	0.95	0.73	0.60	0.09
Random Forest	В	Normal Value	0.73(0.73- 0.73)	0.10	44400	62348	0.18	0.96	0.75	0.62	0.09
Random Forest	С	Normal Value	0.73(0.73- 0.74)	0.09	46774	59974	0.17	0.96	0.77	0.60	0.09
Random Forest	A	No Missing	0.66(0.64- 0.67)	0.26	20159	86589	0.21	0.93	0.40	0.83	0.10
Random Forest	В	No Missing	0.72(0.71- 0.73)	0.14	52201	54547	0.16	0.96	0.78	0.54	0.08
Random Forest	С	No Missing	0.74(0.73- 0.74)	0.21	26193	80555	0.22	0.94	0.55	0.79	0.09

Random			0.66(0.66-	0.45	33868	72880	0.18	0.94	0.57	0.71	0.18
Forest	D	None	0.67)	0.45	33000	72000	0.10	0.54	0.07	0.71	0.10
Neural		Mean	0.71(0.70-	0.47	33169	73579	0.19	0.94	0.60	0.72	0.19
Network	А	Value	0.73)	0.47	55105	13313	0.15	0.54	0.00	0.72	0.13
Neural		Mean	0.74(0.74-	0.49	31171	75577	0.21	0.95	0.63	0.75	0.19
Network	В	Value	0.75)	0.49	511/1	13377	0.21	0.95	0.05	0.75	0.15
Neural		Mean	0.75(0.75-	0.40	42096	64652	0.18	0.96	0.74	0.64	0.10
Network	С	Value	0.76)	0.40	42096	04052	0.18	0.96	0.74	0.64	0.18
Neural		Random	0.74(0.74-	0.50	37930	68818	0.19	0.95	0.68	0.68	0.19
Network	А	Forest	0.74)	0.50	37330	00010	0.15	0.55	0.00	0.00	0.11
Neural		Random	0.75(0.73-	0.54	36554	70194	0.20	0.96	0.71	0.70	0.22
Network	В	Forest	0.76)	0.54	50554	/0154	0.20	0.50	0.71	0.70	0.22
Neural		Random	0.77(0.77-	0.42	42444	64304	0.18	0.96	0.74	0.64	0.18
Network	С	Forest	0.77)	0.42	72777	04504	0.10	0.50	0.74	0.04	0.10
Neural		Extra Trees	0.75(0.75-	0.42	44585	62163	0.18	0.96	0.76	0.62	0.18
Network	А	Regression	0.76)	0.42	44505	02105	0.18	0.90	0.70	0.02	0.10
Neural		Extra Trees	0.76(0.75-	0.46	40067	66681	0.20	0.96	0.75	0.67	0.19
Network	В	Regression	0.76)	0.40	40007	00001	0.20	0.50	0.75	0.07	0.1.
Neural		Extra Trees	0.77(0.77-	0.48	39088	67660	0.19	0.96	0.71	0.67	0.20
Network	С	Regression	0.77)	0.40	55000	07000	0.15	0.50	0.71	0.07	0.20
Neural		Ridge	0.73(0.73-	0.49	43129	63619	0.18	0.96	0.74	0.63	0.22
Network	А	Regression	0.74)	0.73	73123	03013	0.10	0.90	0.74	0.05	0.2.
Neural		Ridge	0.75(0.74-	0.44	39388	67360	0.20	0.96	0.74	0.67	0.18
Network	В	Regression	0.76)	0.44	32300	0/500	0.20	0.90	0.74	0.07	0.10
Neural		Ridge	0.78(0.78-	0.53	38048	68700	0.19	0.95	0.68	0.68	0.22
Network	С	Regression	0.78)	0.55	50040	00700	0.15	0.55	0.00	0.00	0.21

 For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Page	35	of	34
------	----	----	----

 BMJ Open

Neural Network	А	Normal Value	0.72(0.71- 0.72)	0.45	33193	73555	0.19	0.95	0.61	0.72	0.18
Neural Network	В	Normal Value	0.74(0.73- 0.75)	0.50	37560	69188	0.20	0.96	0.71	0.69	0.20
Neural Network	С	Normal Value	0.76(0.76- 0.76)	0.41	42187	64561	0.18	0.96	0.74	0.64	0.18
Neural Network	А	No Missing	0.47(0.45- 0.49)	0.79	2628	104120	0.14	0.90	0.04	0.98	0.44
Neural Network	В	No Missing	0.70(0.68- 0.72)	0.69	23720	83028	0.23	0.94	0.52	0.81	0.31
Neural Network	С	No Missing	0.73(0.72- 0.73)	0.57	51077	55671	0.15	0.95	0.74	0.55	0.29
Neural Network	D	None	0.74(0.73- 0.74)	0.50	36925	69823	0.19	0.95	0.67	0.69	0.21
										0.69	