

Medical University of South Carolina

MEDICA

MUSC Theses and Dissertations

2017

Improved Methods for Modeling High Dimensional Binary Features Data with Applications for Assessing Disease Burden from Diagnostic History and for Dealing with Missing Covariates in Administrative Health Records

Ralph C. Ward

Medical University of South Carolina

Follow this and additional works at: <https://medica-musc.researchcommons.org/theses>

Recommended Citation

Ward, Ralph C., "Improved Methods for Modeling High Dimensional Binary Features Data with Applications for Assessing Disease Burden from Diagnostic History and for Dealing with Missing Covariates in Administrative Health Records" (2017). *MUSC Theses and Dissertations*. 383.
<https://medica-musc.researchcommons.org/theses/383>

This Dissertation is brought to you for free and open access by MEDICA. It has been accepted for inclusion in MUSC Theses and Dissertations by an authorized administrator of MEDICA. For more information, please contact medica@musc.edu.

Improved Methods for Modeling High Dimensional Binary Features Data with
Applications for Assessing Disease Burden from Diagnostic History and for Dealing with
Missing Covariates in Administrative Health Records

by

Ralph C. Ward

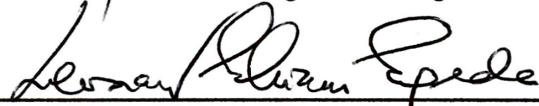
A dissertation submitted to the faculty of the Medical University of South
Carolina in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the College of Graduate Studies.

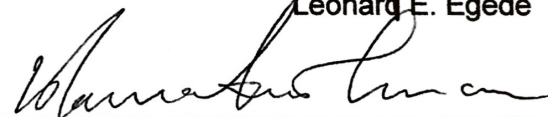
Department of Public Health Sciences

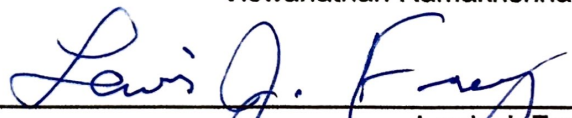
2017

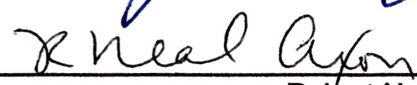
Approved by:
Chairman, Advisory Committee


Mulugeta Gebregziabher


Leonard E. Egede


Viswanathan Ramakrishnan


Lewis J. Frey


Robert Neal Axon

Acknowledgments

I owe a huge debt to many people who enthusiastically supported my unorthodox and eccentric decision to pursue a Ph.D. in middle age. When I first visited MUSC in 2011 to explore the idea, Dr. Viswanathan Ramakrishnan and June Watson were extremely supportive, as they've remained for the past six years. My dissertation mentor, Dr. Mulugeta Gebregziabher, was equally enthusiastic, and his guidance, expertise, and humor were instrumental to my success. The members of my dissertation committee patiently gave valuable hours to help me gradually develop my approaches. Drs. Elizabeth Hill and Bethany Wolf also provided important guidance early in this process. Finally, Dr. Leonard Egede, who founded the VA Health Equity and Rural Outreach Innovation Center, had a particularly strong role in my work since his organization funded this research. As a Veteran, I was grateful to be able to support work that I hope may benefit fellow Veterans.

My parents instilled an appreciation for academic work and life-long learning. My family and close friends cheerfully tolerated and supported this work, for which I'm grateful and fortunate: Christina, Douglas, Emily, and Michael.

Table of Contents

List of Tables	iii
List of Figures.....	iv
1. Abstract	v
2. Introduction.....	1
2.1 Motivation	1
2.2 Specific Aims.....	2
2.3 Background	4
2.4 Significance.....	17
3. First manuscript: Comparison of Statistical and Machine Learning Methods for Developing Improved Comorbidity Models Based on the ICD System.....	19
3.1 Introduction	19
3.2 Study Design and Methods.....	22
3.3. Results	34
3.4 Discussion	43
4. Second Manuscript: Improved Comorbidity Summary Score for Measuring Disease Burden and Predicting Outcomes with Applications to Three National Cohorts	46
4.1. Introduction	46
4.2. Study Design and Methods.....	50
4.3. Results	57
4.4. Discussion.....	61
5. Third Manuscript: Comprehensive Comparison of Machine Learning and Model- Based Multiple Imputation Methods with Competing Sensitivity Analyses for Non- Random Missingness.....	69
5.1. Introduction	69
5.2. Methods	73
5.3 Results	84
5.4 Discussion.....	92
6. Summary and Conclusions	96
6.1 Summary.....	96
6.2 Discussion and Conclusions.....	97
6.3 Limitations	102
6.4 Future Work	103

Table of Contents (continued)

Appendix A: Examples of R and SAS Programs Developed to Support Research Aims	107
References Cited	121

List of Tables

Table 1: Demographic characteristics for the DM and TBI cohorts	35
Table 2: Model performance statistics for Phase 1 models	37
Table 3: Model performance statistics for phase 2 models	40
Table 4: Summary of comorbidities from phase 2 results.....	42
Table 5: Demographic information for the CKD, DM and TBI cohorts.	58
Table 6: Model performance statistics comparison for comorbidity score	60
Table 7: ICD-9-CM conditions that form the summary score.....	65
Table 8: Demographic characteristics for the DM and TBI cohorts	85
Table 9: Comparison of original and updated race-ethnicity distributions.....	86
Table 10: Odds ratios for association between missing race-ethnicity and other covariates.	92
Table 11: Traumatic brain injury (TBI) cohort sensitivity analyses results	93

List of Figures

Figure 1: Example hierarchy for the ICD-9-CM system.....	5
Figure 2: Establishing five year mortality and ICD-9 collection periods	11
Figure 3: Prevalence of the 31 Elixhauser-Quan comorbidities.....	36
Figure 4: Model performance statistics (DM cohort) phase 1	38
Figure 5: Model performance statistics (TBI cohort) phase 1	39
Figure 6: Model performance statistics phase 2.....	41
Figure 7: Overview of summary score development.	52
Figure 8: Model performance statistics summary score	59
Figure 9: Simulation results under MAR	87
Figure 10: Simulation results under MNAR	88
Figure 11: Simulation results: coverage probability.....	89
Figure 12: Simulation results for MNAR sensitivity adjustment:	90

RALPH C. WARD. Improved Methods for Modeling High Dimensional Binary Features Data with Applications for Assessing Disease Burden from Diagnostic History and for Dealing with Missing Covariates in Administrative Health Records. (Under the direction of Mulugeta Gebregziabher).

1. Abstract

Healthcare outcomes research based on administrative data is frequently hindered by two important challenges: (1) accurate adjustment for disease burden and (2) effective management of missing data in key variables. Standard approaches exist for both problems, but these may contribute to biased results. For example, several well-established summary measures are used to adjust for disease burden, often without consideration for whether other methods could perform this task more accurately. Similarly, observations with missing values are often arbitrarily excluded, or the values are imputed without regard for the involved assumptions. Despite recent substantial gains in computing power, statistical approaches and machine learning methods, no comprehensive effort has been made to develop an improved comorbidity index based on predictive performance comparisons of competing approaches. Similarly, recently developed machine learning approaches have shown promise in addressing missing data problems, but these have not been compared with parametric methods via a rigorous simulation study using large-dimensional data with the complete range of missingness types. This makes it difficult to assess the relative merits of each procedure.

This work accomplished three broad aims: (1) Improved models for summarizing disease burden were developed by comparing the predictive performance of a wide

variety of statistical and machine learning methods. (2) A new comorbidity summary score for predicting five-year mortality was developed. (3) A comprehensive comparison of machine learning and model-based multiple imputation methods was completed, both in simulations and through an application to real data. Several sensitivity analyses were also examined for variables with missing not at random (MNAR) missingness.

This work successfully demonstrated several new approaches for summarizing disease burden. Each of the competing disease burden models in the first aim and the summary score from the second aim had superior predictive performance when compared to the Elixhauser index, a commonly-used summary measure. This research also led to new applications for applying machine learning methods within the multiple imputation with chained equations (MICE) framework. Additionally, several MNAR sensitivity methods were adapted and applied to demonstrate that unbiased inference under MNAR may not be possible in some situations, even when the missingness mechanism is fully understood.

2. Introduction

2.1 Motivation

The motivation for this research came from my work in the Veteran's Health Administration's (VHA) Health Services Research and Development (HSR&D) Center of Innovation (COIN) for Health Equity and Rural Outreach, located in Charleston, South Carolina. This group works to reduce disparities in healthcare access and outcomes between Veterans due to racial, ethnic, geographic, or gender-based differences. Much of this research involves observational studies based on VHA and Medicare administrative healthcare datasets, which typically involve millions of patients, each of whom may have thousands of observations involving demographic information, diagnostic and procedure codes, laboratory results, pharmacy records, text notes, and cost data. Most studies are forced to deal with two key challenges:

- 1) Models examining differences between groups must accurately account for each patient's disease burden by summarizing information contained in diagnostic codes, for which there are thousands of unique values.
- 2) Many patients are missing data in key variables that are essential for making any valid inference concerning disparities, such as the race/ethnicity variable. Further compounding the challenge, the pattern for such missingness is often not random.

Standard approaches exist for both problems, but these may risk contributing to biased results. For example, investigators often use the Charlson or Elixhauser Comorbidity indices [1, 2] to summarize a patient's disease burden from the information collected from thousands of covariates, but they may do this without regard for the assumptions and limitations associated with these measures. When dealing with missing data, some investigators exclude observations with missing values and only consider complete cases, while others impute values using models that are based on missing-at-random (MAR) assumptions. Either option could lead to bias, particularly since missingness patterns for some important variables in VHA data likely violate this crucial MAR assumption [3 – 6] .

In developing better approaches for these challenges, it was important to consider the full range of available methods, and to consider whether approaches that combined the strengths of several methods might produce superior results. For example, substantial advances in computing power, statistical and machine learning methods since the Elixhauser index's development in 1998 could support the development of an improved summary measure, perhaps one based on the combined predictions of several methods. Similarly, statistical and machine learning methods each bring different strengths to the missing data problem.

2.2 Specific Aims

2.2.1 Aim 1

Using two large Veteran's Health Administration cohorts involving diabetes and traumatic brain injury, develop improved models for summarizing disease burden from

large-dimension binary diagnostic features data by training and validating models based on a wide variety of statistical and machine learning methods for variable selection and dimension reduction, including a model based on the pooled predictions of the other models. Compare each method's predictive performance with existing scores using AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and net reclassification improvement statistics for events and non-events. Include methods from the following broad categories:

- a. Generalized linear model and regularized regression approaches:
 - (i) Model-Averaged Regression Coefficients (MARC)
 - (ii) Probability Based Features (PBF).
 - (iii) Penalized generalized linear model (elastic-net)
- b. Machine learning methods:
 - (i) Association Rules Analysis (unsupervised method)
 - (ii) Random Forest (supervised method)
- c. Bayesian methods (includes machine learning approaches)
 - (i) Naïve Bayes variable selection (Multi-morbidity Index)
 - (ii) Bayesian Additive Regression Trees

Compare model performance for several mortality outcomes, by applying several methods of establishing baseline comorbidities, and by validating models on both single-disease populations and combined populations.

2.2.2 Aim 2:

Develop a new comorbidity summary score for predicting five-year mortality based on variable importance measures from the top-performing models in the first aim. Train and validate these models using three large VA cohorts with diabetes (DM), chronic kidney disease (CKD), or a history of traumatic brain injury (TBI). Compare the score's

performance to the Elixhauser-Quan index using AUC, sensitivity, specificity, Brier Index, and net reclassification index statistics. Determine if the new score provides any population insights beyond those provided by the existing Elixhauser-Quan index.

2.2.3 Aim 3

Compare machine learning and model-based multiple imputation methods for dealing with missing covariate data under missing at random (MAR) and missing not at random (MNAR) scenarios. For MNAR situations, also examine sensitivity analysis approaches to determine whether unbiased imputation is possible in typical missing data scenarios seen in VA research. Evaluate imputation performance using simulations and by application to VA traumatic brain injury data using relative bias, root mean squared error, efficiency, and coverage probability statistics.

2.2.4 Aim 4:

Publish the R and SAS program code used in each aim on GitHub, along with a simulated dataset that can be used to demonstrate its function.

2.3 Background

2.3.1 Diagnostic code system

In administrative healthcare data, comorbidity information can be found in numerous forms, including physical exam notes, laboratory results and pharmacy records, but this work is focused on that information encoded by the International Classification of Diseases, Clinical Modification (ICD-9-CM or ICD-10-CM), or by a

similar system. These variables consist of 5-digit hierarchical codes, where codes sharing the first three or first four digits are likely to involve related diseases. This hierarchy creates a correlation structure within these data, yet the methods commonly used to model disease burden in ICD-CM data do not account for this structure; nor do they attempt to account for any unidentified interactions. Figure 1 shows the hierarchy for hypertensive chronic kidney disease within the ICD-9-CM system. All codes associated with this condition share the first three digits, while the fourth digit in this example indicates whether the disease is benign, malignant, or unspecified. The fifth digit further classifies the disease.

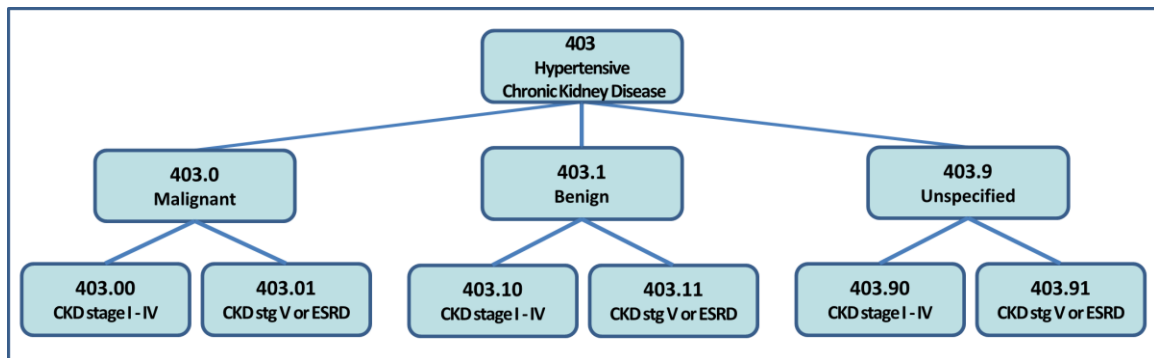


Figure 1: Example hierarchy for the ICD-9-CM system. CKD is chronic kidney disease, and ESRD is end stage renal disease.

2.3.2 Existing comorbidity summary measures

There are several well-known comorbidity summary measures based on the ICD system. The Charlson Comorbidity Index is a score based on the sum of seventeen weighted comorbidities [1]. Deyo et al., Romano et al., Quan et al. and others developed closely-related indices; the newer versions were based on the ICD-9-CM system and thus could be directly applied to administrative databases [7 – 9]. Elixhauser et al.

developed a less parsimonious ICD-9-CM approach for predicting in-hospital mortality that defined 31 comorbidities [2]. Elixhauser excluded numerous conditions such as those related to the primary diagnosis, acute complications related to treatment, or those considered to be unimportant. The Elixhauser index was shown in several studies to be more effective than the Charlson Index in predicting in-hospital mortality and one-year mortality [2, 9, 10] and today remains one of the most commonly used comorbidity indices based on the ICD system. Quan et al. developed enhanced indices that corrected inconsistencies in earlier algorithms and provided better accounting for the ICD taxonomy, which can frequently lead to the same condition being coded in multiple ways.

Van Walraven et al. [11] sought to derive a single score to represent the 31 independent Elixhauser comorbidities, such that it might be easier to develop more parsimonious models, particularly for small populations. This approach produced weights for each comorbidity based on the relative magnitude of predicted coefficients from a multivariate logistic model. Though the authors concluded that neither the summary score nor the original Elixhauser index was effective in predicting in-hospital mortality, they demonstrated their score's predictive ability was as effective as the original Elixhauser index in adjusting for comorbidities based on the comparison of AUC statistics in a dataset of approximately 345,000 hospital admissions.

Quan's enhanced version of the ICD-9-CM Elixhauser index is used throughout the first two aims as the primary basis for comparison since it was shown to have superior predictive performance over the earlier versions of the Elixhauser index [9], and since the van Walraven score was shown to offer no additional advantage. Further

reference to the 'Elixhauser Index' throughout this work thus refers to the Quan enhanced version of the ICD-9-CM Elixhauser Index unless otherwise indicated.

Alemi et al. developed an ICD summary score called the multi-morbidity index based on an application of the Naïve Bayes classification model [12]. This index was applied in several large Veteran's Administration populations [13] to predict mortality within 6 or 12 months, and the authors compared prediction performance against models based on the Quan variant of the Charlson index and the van Walraven variant of the Elixhauser index. The AUC for the multi-morbidity index predicting 6-month mortality was 0.784, compared with values of 0.652 and 0.639 for the Quan-Charlson and van Walraven Elixhauser measures. Although this represents a substantial improvement, the authors do not demonstrate whether the Naïve Bayes approach was the best for binary ICD data, or whether other statistical or machine learning approaches might produce superior results.

2.3.3 Choice of classification models

The first two aims both involve problems of classification. In the first aim, for mortality outcome (y_1, \dots, y_n) and binary ICD-9 predictors (x_1, \dots, x_p) , the goal is to find an unknown function capable of predicting the outcome: $\mathbf{y} = f(\mathbf{x})$. In the second aim, the challenge is similar, except that the ICD-9 binary predictor matrix is replaced by a single summary score for each patient. Existing comorbidity measures such as the Elixhauser or Charlson indices were developed using traditional statistical methods (logistic regression and Cox proportional hazards models) with input from clinicians for decisions on whether to include or exclude various conditions [1, 2]. In order to produce models

with improved classification performance, numerous approaches were considered, including statistical models, machine learning algorithms, and Bayesian methods which incorporated both statistical and machine learning elements. Although there were dozens of methods to consider (see Hastie et al [14]), the intent was to adapt and test as many as feasible, with the goal for finding those with the best classification performance in ICD-9 data, and with the additional goal for finding a collection of methods which succeeded due to dissimilar strengths. For example:

- 1) Some statistical models may succeed based on their ability to account for the correlation structure in ICD-9 data. These data are characterized by hundreds of binary features, many of which are sparse, and many are correlated with other features. This correlation could be due to the hierarchical structure imposed by the ICD system; in other cases it could be due to associations between disease conditions not found within the same hierarchy.
- 2) Machine learning methods may succeed due to their ability to automatically account for unknown interactions and non-linear relationships between predictors [15 – 18].
- 3) Methods based on an ensemble of models may be more successful. Dietterich [19] provided a justification for the observation that ensembles of accurate and diverse classifiers often perform better than the individual models. While his work helps to explain the success of several machine learning methods, it also justifies a model based on the pooled predictions of the successful statistical and machine learning methods from the first aim. Dietterich defined an accurate classifier as one with an error rate lower than that based on random guesses;

diverse classifiers are those with different error rates for the same data. He provided three reasons why an ensemble often provides better results than the individual classifiers [19]. First, given a hypothesis space, H , each classifier provides a hypothesis, with errors associated with the model's inherent characteristics and with the amount of training data provided. When the votes of many classifiers are combined, the overall accuracy will likely improve if the classifiers are truly diverse. Next, models based on searches over the hypothesis space may become fixed on local optima, and an ensemble of models with different search paths will likely provide a better overall solution. Finally, although H theoretically contains all possible hypotheses, its size is in practice limited by the training data's dimensions such that the true classification function might be excluded from H . When the results of numerous models are combined, perhaps in a weighted sum, it may be possible to expand the hypothesis space such that the true classification function is found.

2.3.4 Establishing baseline disease burden

Each aim involves observational data in which patients were included at the start of the study if they met diagnostic criteria for the primary disease (diabetes, chronic kidney disease, or traumatic brain injury), and additional patients entered each year of the study as they first met the same criteria. Patients were followed until death or the end of the study. There was no "dropout" category: patients who had no in-patient or out-patient visits in a given year were assumed to be alive unless a date of death was found. In many cases an exact diagnosis date was unknown. For example, in TBI patients, the

original injury may have occurred in combat, and the injury date would likely only be found in the patient's Department of Defense medical record, which was not available in this study. Similarly in other cohorts, the original diagnosis may be recorded by another healthcare system. Given these limitations, mortality outcomes were defined based on how many years the patient lived after entering the study. Each patient's set of unique ICD-9 codes were collected from the earliest entry in the patient's VHA record until an appropriate cutoff point before death, or until the study's end, as applicable. For five year mortality, this cutoff was arbitrarily set as follows:

- 1) If the patient died within five years of entering the study, the cutoff was set at one year prior to death. This excluded codes for conditions that typically occur just prior to death, such as palliative care; these conditions are highly associated with the outcome but are of less use in making long-term predictions.
- 2) If the patient died after being in the study for more than five years, the cutoff for ICD code collection was set at five years after study entry.
- 3) If the patient did not die during the study, the cutoff was set at the study's end date.

Figure 2 illustrates these limits for two patients (A and B). Regions shaded in red are ICD-9 code collection periods for patients A and B. Patient A entered the VA system in 1985 and entered the study in 2000; patient B entered the VA system in 2003 and entered the study at the same time. Patient A died within 5 years of the study's start date, while patient B was still alive 5 years after entering the study and was recorded as "alive" in the five-year mortality variable. Patient B's ICD code collection stopped five years after he or she entered the study. The two patients may have substantially

different total numbers of ICD codes, and the challenge for the models was similar to asking, “given everything in the record up until this point, what is the probability the patient actually died within five years of entering the study?” For the models to provide good predictions, the presence of more or less information for a given patient should not lead to bias in either direction. The comparison models based on the Elixhauser index faced the same challenge. While there are several possible ways to establish the limits described here, this method was found to produce reasonable results.

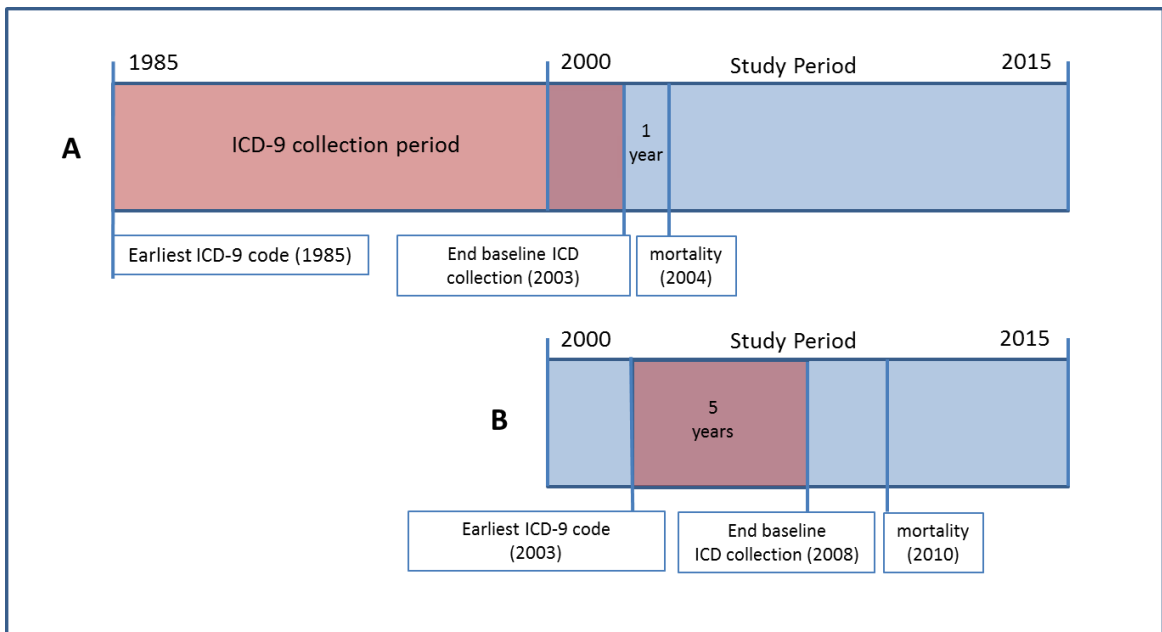


Figure 2: Establishing five year mortality and ICD-9 collection periods for a study running between 2000-2015.

2.3.6 Evaluation of model performance

2.3.6.1 Evaluation of classification models (Aims 1 and 2)

Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were all considered when comparing models. The net reclassification improvement (NRI) statistic

for events (patients who died) and non-events (patients who lived) was also considered [20]. NRI statistics for the Elixhauser models are not provided since they form the reference. While NRI statistics have been widely adopted, Pepe et al. [21] showed they should be used with caution. In particular, the authors demonstrated that positive NRI results could be achieved in some situations where the new model involved an added variable with no predictive value, possibly due to poorly fitting risk models. The AUC and related ROC statistics provided reliable results in these situations. Primary emphasis was thus placed on the AUC and related statistics; in particular, the NRI result was not claimed as evidence for prediction performance improvement unless similar gains were seen in the AUC. Finally, the Brier Score provided a measure of misclassification error or mean-squared error for binary outcomes [22].

2.3.6.2 Evaluation of multiple imputation models (Aim 3)

Imputation methods were compared using the following statistics:

- 1) Relative bias: $(\hat{\beta} - \hat{\beta}^o) / \hat{\beta}$, where $\hat{\beta}$ and $\hat{\beta}^o$ are the generalized linear model parameter estimates based on the imputed data and the full dataset of complete cases, respectively.
- 2) Efficiency: $\text{var}(\hat{\beta}) / \text{var}(\hat{\beta}^o)$
- 3) Root mean square error: $\sqrt{(\hat{\beta} - \hat{\beta}^o)^2 + \hat{\sigma}^2}$, where $\hat{\sigma}^2$ is the estimated variance of the parameter estimate from the model based on imputed data.
- 4) Coverage probability: the probability based on 1000 bootstrapped iterations that the 95% confidence interval for the parameter estimate contains $\hat{\beta}^o$.

2.3.7 Missingness in VHA data

Missing data in key VHA variables such as race/ethnicity poses a substantial problem for investigators involved in healthcare inequities. Further, several investigators have reported missing race information in VHA or Medicare data may not be missing at random [4 – 6]. Depending on the timeframe being studied, the level of missingness may be substantial. Stroupe et al. [23] reported that 48% of VHA patient records had missing race-ethnicity information in 2004, but this value had been reduced to 15% by 2012 [3] due to concerted efforts to collect this information and due to a 2003 requirement for recording self-reported race-ethnicity rather than observer-reported values [24]. Stroupe et al. [23] demonstrated that further improvements were possible by merging VHA data with Medicare data; in the author's experience with several VHA cohorts followed through 2012 or later, the missing race fraction can now be reduced to below 5%. However, the missing race-ethnicity problem is far from solved: even at these lower levels, if the data are believed to be missing due to non-random processes, investigators must still be concerned whether unbiased results were achieved. Further, studies involving patients who were not followed in recent years will likely still face substantial missing data problems.

2.3.8 Existing multiple imputation methods

In past years, researchers often dealt with the missing data problem by simply conducting complete-case analysis, though this strategy could lead to biased results unless the data were missing completely at random. More recently, steps to attempt to assess the pattern of missingness and methods to help achieve unbiased results are commonly seen. Numerous parametric imputation methods exist for handling data with missing completely at random (MCAR) or missing at random (MAR) patterns; multiple imputation with by chained equations (MICE) is one commonly used approach due to its ability to handle multiple imputation for mixed data types [25, 26]. MICE imputes missing values from separate conditional distributions for each variable with missing values, but has been criticized for lacking a theoretical basis [27], and for requiring the investigator to have advance knowledge of non-linear relationships or collinearities between predictors [17]. Other researchers have concluded that machine learning methods can automatically handle interactions and other concerns while also producing inference estimates with narrower confidence limits and with more computational efficiency. The random forest algorithm has been applied in several multiple imputation research efforts, and involves bootstrap aggregation of numerous independent decision trees, and can account for complex interactions and collinearities between predictors more readily than many parametric methods, while the ensemble voting of independent trees naturally lends itself to an efficient imputation process [28]. For example Stekhoven et al. [16] claim their multiple imputation approach (missForest) based on the random forest method was superior to traditional statistical methods including MICE, based on improved misclassification error rates or normalized root mean squared errors. Jerez et

al. [29] provided a similar conclusion based on a comparison of machine learning and statistical imputation methods. Other researchers have incorporated machine learning methods within an existing statistical method. For example, Shah et al. [17] incorporated random forest as the multiple imputation method within the existing MICE method and showed the new approach had a superior ability to handle nonlinear relationships and collinearities.

2.3.9 Evaluating missing not at random (MNAR) situations

Though the multiple imputation methods described above are capable of producing unbiased results under MCAR and MAR, such results are far less likely under MMAR. As Verbeke et al. [30] discuss, it is possible to construct models based on MNAR assumptions, but these assumptions are not testable since their support is not contained in the data. Further, Molenburghs et al. [31] demonstrated that it is not possible to empirically distinguish between MNAR and MAR situations from the data alone because for every MNAR model, it is possible to build an MAR model with the same fit. The most common approach given these circumstances is to conduct sensitivity analysis on MAR models to examine their stability when MNAR assumptions are introduced [32, 33] Though numerous approaches are possible, two general types of sensitivity analyses are most common; these are based on pattern mixture models [32 – 34] and selection models [35].

A pattern mixture model assumes that a number of missingness patterns may exist, each with a separate joint distribution for the partially and fully observed variables. For patients $i = 1, \dots, n$ and covariates Y_{1i} and Y_{2i} , assume Y_{1i} has missing values with

indicator R_i , such that $R_i = 0$ when Y_{1i} is missing and $R_i = 1$ otherwise. Under MNAR the joint distribution $f(Y_{1i}, Y_{2i}, R_i)$ is factored as $f(Y_{1i}, Y_{2i} / R_i) f(R_i)$, where the joint distribution of the partially and fully observed variables is conditional on the partially observed variable. Since the MNAR distribution cannot be determined from the observed data, Carpenter and Kenward [32] suggest starting from the MAR scenario and then adjusting the model using MNAR assumptions in order to examine whether the MAR assumption is sensitive to such changes.

A selection model, on the other hand, factors the joint distribution $f(Y_{1i}, Y_{2i}, R_i)$ differently; now the focus is on the mechanism behind the MNAR process:

$f(Y_{1i}, Y_{2i}, R_i) = f(R_i / Y_{1i}, Y_{2i}) f(Y_{1i}, Y_{2i})$. Numerous methods are based on this factorization; in the third aim, a weighting approach is applied [36].

2.3.10 Resampling Methods (Aims 1 - 3)

Resampling methods were applied for several reasons:

- 1) Some methods, including Bayesian additive regression trees and random forest could not be run in a reasonable amount of time on large datasets involving millions of patients without resorting to a parallel computing environment. Instead, a resampling approach was used to generate model performance estimates. For example, in the first aim, 1000 smaller test and training datasets of 5000 observations each were generated by randomly sampling the full datasets with replacement. Performance statistics were collected for each validation run and the

overall mean and 95% confidence intervals generated by 1000 iterations were used to compare the models' relative performance.

- 2) In simulations, a resampling approach was used to generate large numbers of independent training and validation datasets from actual VHA data rather than relying on fully-generated data. This helped to ensure that the complex structures and associations found in real patient observations were also present in synthetic datasets. This was particularly important due to the complex correlation structure in ICD-9 data. As demonstrated by Marshall et al. [37] and Gebregziabher et al. [38] this approach is reasonable when the original dataset is large enough to help assure independence between samples.

When applying resampling methods, steps were taken to ensure full independence between training and validation datasets. In the first aim, training and validation datasets were generated in pairs during each iteration, with steps taken to ensure no observations were common to the two sets during the bootstrapping process. In the second aim, 1000 training data sets were used to determine variable importance measures, which were then used to determine the comorbidity score. The score, in turn, was tested on 1000 validation data sets. Because the validation step took place after all of the training datasets had been analyzed, it was necessary to randomly partition the full dataset such that training data was drawn from one subset, and validation data from the other. This ensured that validation data had not been used in model development.

2.4 Significance

This research made new contributions in the following areas:

- 1) Although other research has compared traditional statistical methods and machine learning approaches in the development of predictive models for specific disease conditions [39 – 44], to the best of the author's knowledge, this is the first effort to conduct a detailed application of such methods in the development of improved ICD-based disease burden models (aim 1) and an improved ICD-based summary score (aim 2). In the first aim, the best models (Bayesian additive regression trees, random forest, elastic-net and the pooled model) consistently had better predictive performance when compared with the Elixhauser index. Similarly in the second aim, the comorbidity summary score for predicting five-year mortality had stronger predictive performance than the widely-used Elixhauser index.
- 2) This research provided a comprehensive comparison of multiple imputation methods under both MAR and MNAR conditions, and in particular, developed new applications for applying machine learning methods within the multiple imputation with chained equations (MICE) framework. Additionally, several MNAR sensitivity methods were adapted and applied, both in simulations and in actual data, to demonstrate that unbiased inference may not be possible in some MNAR scenarios, even when the missingness mechanism is fully understood. This result has direct implications for VHA research involving missing race/ethnicity data.

3. First manuscript: Comparison of Statistical and Machine Learning Methods for Developing Improved Comorbidity Models Based on the ICD System

3.1 Introduction

When conducting healthcare outcomes research, accounting for disease burden is essential for reducing the potential for bias in estimating the association between outcomes and risk factors. For example, researchers designing studies to examine disparities between racial and ethnic groups with diabetes must first account for each patient's other diseases and conditions; otherwise, the study is not likely to produce meaningful results. Since this research frequently involves administrative healthcare databases or electronic health records, this effort will become increasingly important as the availability and quantity of such data continues to rapidly expand.

In administrative healthcare data, comorbidity information is found in numerous forms, including physical exam notes, laboratory results and pharmacy records, but this paper is concerned with that comorbidity information encoded by the International Classification of Diseases, Clinical Modification (ICD-9-CM or ICD-10-CM), or by a similar system. These variables consist of hierarchical codes. For example, in the ICD-9-system, codes sharing the first three or first four digits are likely to involve related diseases. This hierarchy creates a correlation structure, yet the methods commonly used to model disease burden in ICD-CM data do not account for this structure. Nor do they

attempt to account for any unidentified complex interactions and frequently do not consider disease severity.

There are several well-known comorbidity summary indices based on the ICD system. The Charlson Comorbidity Index [1] is a single score based on the sum of 17 weighted comorbidities. Deyo et al. [45], Romano et al. [8], Quan et al. [9] and others developed closely related indices based on the same 17 comorbidities. In contrast to the Charlson's single summary score, Elixhauser et al. [2] developed a more complex index that consisted of 31 distinct comorbidities. Because the outcome was in-hospital mortality, Elixhauser excluded conditions related to the primary diagnosis, acute complications related to treatment, or those considered unimportant. The Elixhauser index was shown in several studies to be more effective than the Charlson Index for predicting in-hospital mortality and one-year mortality and today remains one of the most commonly used comorbidity indices based on the ICD system [2, 9,10]. Quan et al. [9] developed enhanced indices that corrected inconsistencies in earlier algorithms and added improved accounting for the ICD taxonomy, where the same condition might be coded in several ways. Quan's enhanced version of the ICD-9-CM Elixhauser index is used as the basis for comparison here since it was shown to have superior predictive performance over the earlier versions [9].

Kheirbek et al. [12] developed an ICD summary score called the multi-morbidity index based on an application of the Naïve Bayes classification model. This index has been applied in several large Veteran's Administration populations to predict mortality within 6 or 12 months [13]. The multi-morbidity index's prediction performance was shown to be superior against models based on the Quan variant of the Charlson index

and the van Walraven variant of the Elixhauser index [11]. However, the authors only considered the naïve Bayes approach, and did not demonstrate whether other statistical or machine learning methods might produce superior results. Similarly, Siddique et al. [46] relied on a single method (classification trees) to develop an ICD-9-CM based algorithm for predicting which patients had lower gastrointestinal bleeding.

This work is based on the premise that advances in computing power, machine learning and statistical methods since the Elixhauser Index's introduction in 1998 will support the development of improved ICD-based models with better predictive performance. Seven statistical and machine learning methods for the analysis of high dimensional data with binary ICD-CM predictors are compared. These methods apply various approaches that were not considered in the Elixhauser index's development, including (1) empirically identifying latent features, (2) accounting for the inherent hierarchical structure in ICD-CM data, (3) automatically incorporating complex interactions, and (4) attempting to account for disease severity. Although other research has compared traditional statistical methods and machine learning approaches in the development of predictive models for specific disease conditions [39 – 44], to the best of the author's knowledge, this is the first effort to conduct a detailed comparison of such methods in the development of an improved ICD-CM based comorbidity summary measure.

This research is focused on the ICD-9-CM rather than the ICD-10-CM system because it involves Veteran's Administration data recorded under the ICD-9 system, though the same methods could easily be applied to ICD-10-CM data.

3.2 Study Design and Methods

This study was conducted in two phases. In the first phase, the seven methods described below were applied, each in separate models for two populations. The outcome was mortality within the study's timeframe, and no ICD-9 codes were excluded based on their temporal proximity to the patient's death. Each model's predictions were compared to those from models based on the Elixhauser-Quan comorbidities derived from the same ICD-9 data. Each model was trained using a single-disease population and was validated using other patients drawn from the same population. In the second phase, the top four models based on predictive performance from the first phase were used to examine whether performance varied when the outcome was shifted to five-year mortality instead of death within the study's timeframe. Further, ICD-9 codes recorded within one year of death were excluded since these might provide an unrealistic advantage over the Elixhauser-Quan index. For example, the ICD-9 code for palliative care is strongly associated with death but may not be useful for predicting mortality several years in the future. Finally, in the second phase each model was again trained in a single-disease population, but was now validated on a combined population equally drawn from the traumatic brain injury and diabetes groups in order to examine how well they each performed in a more general setting with a wider range of comorbidities.

3.2.1 Populations

Two national cohorts of U.S. Veterans were used; these had been created for earlier studies by linking numerous Veterans Health Administration patient and administrative databases. The first included 625,903 patients with diabetes mellitus (DM)

based on two or more related ICD-9-CM codes and at least one prescription filled for a medication to treat diabetes [47]. In the original study, Veterans were followed from 2002 until death, loss to follow-up, or until December 2006, and newer data was added to extend the follow-time until December 2012. The second cohort involved 168,125 Veterans diagnosed with traumatic brain injury (TBI) during 2004 and 2005. In the original study, patients were followed from the point of entry until death, loss to follow-up, or until December 2010 [48]; newer data were added to extend the follow-time to December 2014. Both studies were approved by the Medical University of South Carolina Institutional Review Board (IRB) and the Ralph H. Johnson Veterans Affairs Medical Center Research and Development committee.

3.2.2 Patient demographic and clinical covariates

Models termed 'unadjusted' used only on ICD-9 predictors; those termed 'adjusted' also controlled for each patient's demographic and clinical covariates. In both cohorts, the patient's age in years was treated as a continuous variable. Race and ethnicity were categorized as non-Hispanic white, non-Hispanic black, Hispanic, and other / missing. Gender and marital status were treated as binary variables. In the TBI cohort, homeless status was treated as a binary variable. TBI severity was categorized as 'not severe', 'moderately severe', and 'severe'. A binary variable was used to indicate if the TBI injury was related to military service. The patient's location was categorized over the five Veterans Administration (VA) regions. The patient's location by Rural Urban Commuting Area (RUCA) code was categorized as 'urban', 'rural', and 'highly rural'.

Finally, the availability of poly-trauma treatment centers by VA station and by VA integrated service network (VISN) were each included as binary variables.

Prior to analysis, erroneous duplicate formats of ICD-9 codes were identified and corrected to the single correct format. For example, codes 250, 2500, and 25000 all represent the same condition, but each would be treated as separate predictors in machine learning algorithms. Next, the listing of unique ICD-9 codes gathered from all patients was ranked by frequency. The 1000 most frequent codes formed the feature set used in subsequent analyses, and this listing accounted for approximately 90% of all ICD-9 codes recorded among all patients in the respective datasets.

3.2.3 Outcomes

In the first phase, the outcome was death within the study window; in the second phase, the outcome was five-year mortality.

3.2.4 Methods

Each phase was conducted in two parts: (i) first prediction models were developed using training datasets; (ii) each prediction model was then validated using test data. Since computational efficiency was a concern for some methods due to the very large datasets involved, resampling methods were used to generate 1000 smaller test and training datasets of 5000 observations each by randomly sampling the full datasets with replacement. Performance statistics were collected for each validation run, and their mean and 95% confidence intervals were determined over 1000 iterations. As demonstrated by Marshall et al. [37] and Gebregziabher et al. [49], this non-parametric

bootstrapping approach is reasonable for these large datasets, such that independence between numerous samples is reasonably assured.

3.2.4.1 Generalized linear model with penalized maximum likelihood (elastic-net regression):

Several penalized generalized linear models were considered, including ridge regression [50], LASSO regression [51], elastic-net regression [52], and group LASSO regression [53]. Elastic-net regression provided the best predictive performance in the ICD-9-CM datasets. The elastic-net model incorporates both the LASSO and ridge approaches with the addition of parameter α such that the loss function becomes the LASSO model when α is 1, and the ridge model when α is 0. An iterative process showed that $\alpha = 0.5$ provided the best predictive performance. For binary outcome $y \in (-1, 1)$, predictors $x_i : (1, x_{1,i}, \dots, x_{p-1,i})$ and shrinkage parameter λ , the following equation was minimized in order to determine coefficient estimates [14]:

$$\max_{\beta_0, \beta} \left[\sum_{i=1}^n \{y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i})\} - \lambda \sum_{j=1}^p \{\alpha |\beta_j| + (1 - \alpha) \beta_j^2\} \right].$$

The R package *glmnet* [54] was used to determine parameter estimates based on the training data, and then used these estimates to generate predictions in the test data.

3.2.4.2 Model averaged regression coefficients (MARC):

This model is based on adapting a method developed by Glance et al. [55] for their Trauma Mortality Prediction Model. This approach attempted to account for the correlation structure created by the ICD-9-CM hierarchy. The first step involved creating two generalized linear models using a probit link. The first model used the full ICD-9-CM

code as it appears in the data. For the second model, Glance et al. relied on a separate scale for trauma location and severity to group related injuries into higher level ‘bins’. This would not be practical here since the models include every unique disease code rather than a limited set of trauma injuries. Instead, each ICD code was collapsed to its first three digits, thereby combining all information for a given hierarchy of related comorbidities into a single high-level variable. The estimated coefficients for the two models were then combined using an inverse variance weighting approach such that the high-level model coefficients were weighted more when the variance of the corresponding coefficient estimate was lower than that for the full model. Thus, when there was little information about a particular comorbidity (and thus a higher estimated coefficient variance), information from the related comorbidities in the hierarchy was given a stronger weight.

The full model, which relied on the 5 digit ICD-9-CM code, was written:

$$P(\text{death}) = \Phi \left(\gamma_0 + \sum_{i=1}^{1000} \gamma_i x_i + \alpha_1 \text{age} + \alpha_2 \text{gender} + \alpha_3 \text{rural} + \alpha_4 \text{race} + \alpha_5 \text{meds} + \alpha_6 \text{marital} \right),$$

where Φ is the probit link, γ_i and x_i are the coefficient and binary indicator for the i^{th} ICD-9-CM code, and α_i is the coefficient for a given patient covariate.

The high-level model, which collapsed data to the first three digits of the ICD-CM code, was written:

$$P(\text{death}) = \Phi \left(\beta_0 + \sum_{j=1}^J \beta_j z_j + \alpha_1 \text{age} + \alpha_2 \text{gender} + \alpha_3 \text{rural} + \alpha_4 \text{race} + \alpha_5 \text{meds} + \alpha_6 \text{marital} \right),$$

where β_j and z_j are the j^{th} coefficient and binary indicators for the J high level ICD-CM variables created when the ICD-CM codes were collapsed to the highest level. The parameter estimates from the two models were combined using weighted inverse variances to produce a Model Averaged Regression Coefficient (MARC) for each of the top 1000 ICD-CM predictors. That is,

$$MARC_i = \frac{\frac{1}{var(\hat{\gamma}_i)}}{\frac{1}{var(\hat{\gamma}_i)} + \frac{1}{var(\hat{\beta}_j)}} \hat{\gamma}_i + \frac{\frac{1}{var(\hat{\beta}_j)}}{\frac{1}{var(\hat{\gamma}_i)} + \frac{1}{var(\hat{\beta}_j)}} \hat{\beta}_j,$$

where $var(\hat{\beta}_j)$ is a weighted variance of the $\hat{\gamma}_i$'s that map to a specific $\hat{\beta}_j$, where each $\hat{\gamma}_i$'s contribution to the overall variance was weighted by its inverse variance [55]:

$$var(\hat{\beta}_j) = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} N_j W_i [\hat{\gamma}_i - E(\hat{\gamma})]^2,$$

$$\text{where } W_i = \frac{1}{N_j \sum_{k=1}^{N_j} var(\hat{\gamma}_k)} \text{ and } E(\hat{\gamma}) = \sum_{i=1}^{N_j} W_i \hat{\gamma}_i.$$

Finally, new predictions were made using test data and the sum of each patient's MARC values associated with their ICD-9-CM codes. In summary, the MARC values were determined from training data, and were then applied in test data for new patients in the validation model:

$$P(\text{death}) = \Phi(C_0 + C_1(MARCsum) + \alpha_1 age + \alpha_2 gender + \alpha_3 rural + \alpha_4 race + \alpha_5 meds + \alpha_6 marital).$$

3.2.4.3 Naïve Bayes Variable Selection (multi-morbidity index)

Kheirbek et al. [12] developed an ICD summary score called the multi-morbidity index based on an application of the Naïve Bayes classification model. For binary random variables $\mathbf{X} = (X_1, \dots, X_r, \dots, X_d)$ sampled from a population classified by two categories (i or j), the odds of outcome i occurring are:

$$\frac{P(i|x)}{P(j|x)} = \frac{P(x|i)P(i)}{P(x|j)P(j)}.$$

Each random variable is assumed to be independent of the others:

$$\frac{P(i|x)}{P(j|x)} = \frac{P(i) \prod_{r=1}^d P(x_r | i)}{P(j) \prod_{r=1}^d P(x_r | j)}.$$

The posterior probability for outcome i can be easily calculated from the posterior odds above. The assumption of independence among the predictors in \mathbf{X} is questionable, and Hand et al. [56] discuss reasons why this approach is nonetheless often successful. In particular, the authors argue that although the Naïve Bayes approach may produce biased estimates, the variance for such estimates is often lower than seen in less parsimonious models. Further, for classification purposes such bias is not a hindrance as long as it is in the right direction. Kheirbek et al. made several necessary accommodations in order to apply the Naïve Bayes approach to ICD-9 data. For perfectly separated predictors, the posterior odds were arbitrarily defined as $1/n+1$ when all patients died and $n+1$ when all patients survived, where n is the number of patients with a given ICD-9 code. Next, when the number of patients with a given ICD-9 code was small, data from related diagnoses in the same ICD-9 hierarchy were

combined based on the assumption that related conditions have similar associations with the outcome of interest.

3.2.4.4 Association Rules Analysis [14]:

This is an unsupervised machine learning method concerned with finding joint values of predictors (X_1, X_2, \dots, X_p) that appear most often in the data. Because ICD-9-CM data is binary, the support for each X_j is $S = \{0,1\}$, and the goal is to find conjunctive rules based on regions in the X space with a larger probability for joint occurrences:

$$\Pr \left[\bigcap_{j=1}^p (X_j \in s) \right],$$

where s is a single value of the support for X_j . Next, the conjunctive rules are transformed to become:

$$\Pr \left[\bigcap_{k \in K} (Z_k = 1) \right] = \Pr \left[\prod_{k \in K} (Z_k = 1) \right],$$

where Z_k represents a binary dummy variable formed from one level of X_j . The set of predictors in conjunctive rule K is called the item set, and the number of Z_k variables in the set is known as the size. The estimated value for a conjunctive rule is called the support or prevalence T :

$$T(K) = \frac{1}{N} \sum_{i=1}^N \prod_{k \in K} z_{ik},$$

where z_{ik} is the value of Z_k for the i^{th} observation. T thus represents the proportion of observations which contain the conjunctive rule. When the item set K is divided into two parts, such that antecedent A predicts the presence of consequent B , $T(K)$ becomes $T(A \rightarrow B)$. The confidence C for this association could be viewed as finding $P(B|A)$:

$$C(A \rightarrow B) = \frac{T(A \rightarrow B)}{T(A)}.$$

When the predictors which make up A appear in an observation, the confidence value represents the probability that predictor B will also appear. Hastie et al. [14] comment that association rules analysis is very good at finding combinations of variables that appear frequently, but is less good at finding those with lower support. Thus we would not expect to identify a joint occurrence that included at least one rare ICD-9-CM diagnosis, even if this joint occurrence were strongly associated with the outcome.

The R package *arules* was used to implement association rules analysis [57]. Prevalence and confidence thresholds t and c were set to limit the number of rules returned by the algorithm:

$$T(A \rightarrow B) > t \text{ and } C(A \rightarrow B) > c.$$

Because this is an unsupervised method, joint occurrences were identified without regard to the outcome of interest. Each candidate rule was tested for significance against the outcome on a univariate basis using a different set of training data than that used to generate the association rule. Multiple testing was accounted for during this process using the Bonferroni adjustment, such that the critical value for significance ($p=0.05$) was divided by the total number of rules that were tested. Next, LASSO

regularized logistic regression was used to help determine which rules were most important in predicting the outcome. The resulting association rules and their parameter estimates were then used to make predictions for other patients in a test data set.

3.2.4.5 Random Forest (RF) [28]

This is a well-known ensemble method based on classification trees that relies on bootstrap aggregation (or ‘bagging’) to generate a forest of generally uncorrelated trees, where each tree then votes for the predicted outcome. The forest is termed “random” due to the random selection of a pre-specified number of features (or predictor variables) at each node; the feature that leads to the largest improvement in the tree’s classification ability is then used to split the data at that node. The random forest method can identify complex interactions, and was reported to be very competitive with other machine learning methods when compared on the basis of misclassification error [14]. The R package *randomForest* was used to implement this method [58]. A forest was generated using patients in a training dataset, which was then used to make predictions on other patients in test data.

2.4.6 Bayesian Additive Regression Trees (BART) [59]

This is an extension of the supervised tree-based ensemble learning method, but unlike random forest, prior distributions are established for each tree’s decision rules and terminal node parameters, and an MCMC algorithm is used to sample from the posterior distribution for the ensemble of trees. The authors contend their approach provides a substantial degree of regularization such that each tree’s complexity is

reduced. They also claim that the predictive results in some datasets were superior to random forest, neural nets, and regularized regression methods [59]. For a BART model with binary outcome Y , a probit model is used:

$$P[Y = 1 | x] = \Phi[G(x)],$$

where x represents the data and $G(x)$ is a summation of m trees, where the j^{th} tree is designated as $g(x; T_j, M_j)$:

$$G(x) = \sum_{j=1}^m g(x; T_j, M_j).$$

Here, T and M are the tree's decision rules and terminal node parameters, respectively, and each is assumed to have independent and identical prior distributions. The prior for T_j is defined in multiple parts. First, the probability that a given node of depth d is non-terminal is $\alpha(1+d)^{-\beta}$, for $\alpha \in (0,1), \beta \in [0, \infty)$. Values of 0.95 and 2 were selected for α and β , respectively in order to help limit each tree's size. Finally, uniform priors were used to model the splitting variable and splitting rule assignments for interior nodes in each T_j . For M_j , a Gaussian prior distribution is assumed for the mean value μ_{ij} for terminal node i within tree j :

$$\mu_{ij} \sim N(0, \sigma_\mu^2), \text{ where } \sigma_\mu = 3.0 / k\sqrt{m},$$

where m is the number of trees, typically 200, and k is a parameter typically set between 1 and 3. This prior serves to limit the values of $G(x)$ to within $(-3.0, 3.0)$, and thus shrinks $G(x)$ towards 0 and $P(Y=1|x)$ towards 0.5.

The R package *BayesTree* was used to implement BART [60]. This algorithm develops the model based on the training data, and then provides the results of post-convergence samples from the posterior distribution using test data. The mean or median of these samples is then used to provide a prediction for each test set observation. The algorithm also returns the number of times each predictor is used in a decision rule among all trees; this serves as a variable importance measure.

3.2.4.7 Pooled prediction model

An ensemble model was developed based on the combined predictions from all of the models considered. Here each model's test data predictions were used as independent predictors in a logistic model using the same test data. The predictive performance of this combined model was then compared against the other six models.

3.2.4.8 Elixhauser-Quan comparison model

Each of the above approaches was compared against a model based on the 31 independent Elixhauser-Quan comorbidities:

$$P(\text{death}) = \Phi \left(C_0 + \sum_{i=1}^{31} C_i (\text{Elix_comorb}_i) + \alpha_1 \text{age} + \alpha_2 \text{gender} + \alpha_3 \text{rural} + \alpha_4 \text{race} + \alpha_5 \text{medcount} + \alpha_6 \text{marital status} \right),$$

where C_i is the estimated coefficient for the i th comorbidity from the enhanced Elixhauser-Quan index [9].

3.2.4.9 Model Performance Assessment

Area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were all considered when comparing models. Net reclassification improvement (NRI) was also reported for events (patients who died) and non-events (patients who lived) [20]. While NRI statistics have been widely adopted, Pepe et al. [21] showed they should be used with caution. In particular, the authors demonstrated that positive NRI results could be achieved in some situations where the new model involved an added variable with no predictive value, possibly due to poorly fitting risk models. The AUC and related ROC statistics provided reliable results in these situations. Primary emphasis was thus placed on the AUC and related statistics; in particular, a strong positive NRI result was not claimed as evidence for prediction performance improvement unless similar gains were seen in the AUC. Finally, the Brier Score was reported as a measure of misclassification error [22].

3.3. Results

Table 1 provides demographic information for the two populations examined in this study and Figure 3 provides the percentage of each group diagnosed with each of the 31 Elixhauser comorbidities. The DM cohort was older than the TBI group (mean age 73.1 versus 49.9), and had a higher five-year mortality rate (13.3% versus 4.4%). The DM cohort had higher rates of congestive heart failure, peripheral vascular disorders, hypertension, diabetes complications, and renal failure when compared to the TBI group.

Table 1: Demographic characteristics for the Diabetes Mellitus (DM) and Traumatic Brain Injury (TBI) cohorts

Variable	Level	Diabetes (n=625,903)	Traumatic Brain Injury (n=168,125)
Five-year mortality (%)		13.3	4.4
Mean age		73.1	49.9
Gender (%)	male	97.0	93.7
	female	3.0	6.3
Marital status (%)	single	7.0	26.0
	widowed	11.0	4.6
	divorced	21.0	25.9
	married	59.0	42.1
Race/ethnicity (%)	Non-Hispanic white	76.0	55.8
	Non-Hispanic black	15.0	13.0
	Hispanic	5.0	1.9
	Other or missing	4.0	29.2
Homeless (%)		8.0	1.5
Greater than 50% disability (%)	(service-connected)	27.0	23.3

The TBI cohort had substantially higher rates of depression, psychoses, drug abuse, alcohol abuse, liver disease, and neurological disorders.

Table 2 provides a comparison of validation results for the DM and TBI cohorts for the seven models that were compared against the Elixhauser-Quan model in phase 1, and Figures 4 and 5 provide a corresponding graphical comparison of confidence intervals for each statistic based on 1000 iterations. Results labeled “unadjusted” correspond to models in which only ICD-9-CM codes were used as predictors, while “adjusted” models also included patient demographic variables. Overall, the BART, random forest, elastic-net and pooled models had the best predictive performance as seen in their consistently higher mean AUC values and lower Brier scores for unadjusted and adjusted models in both cohorts. The MARC, association rules and multi-morbidity

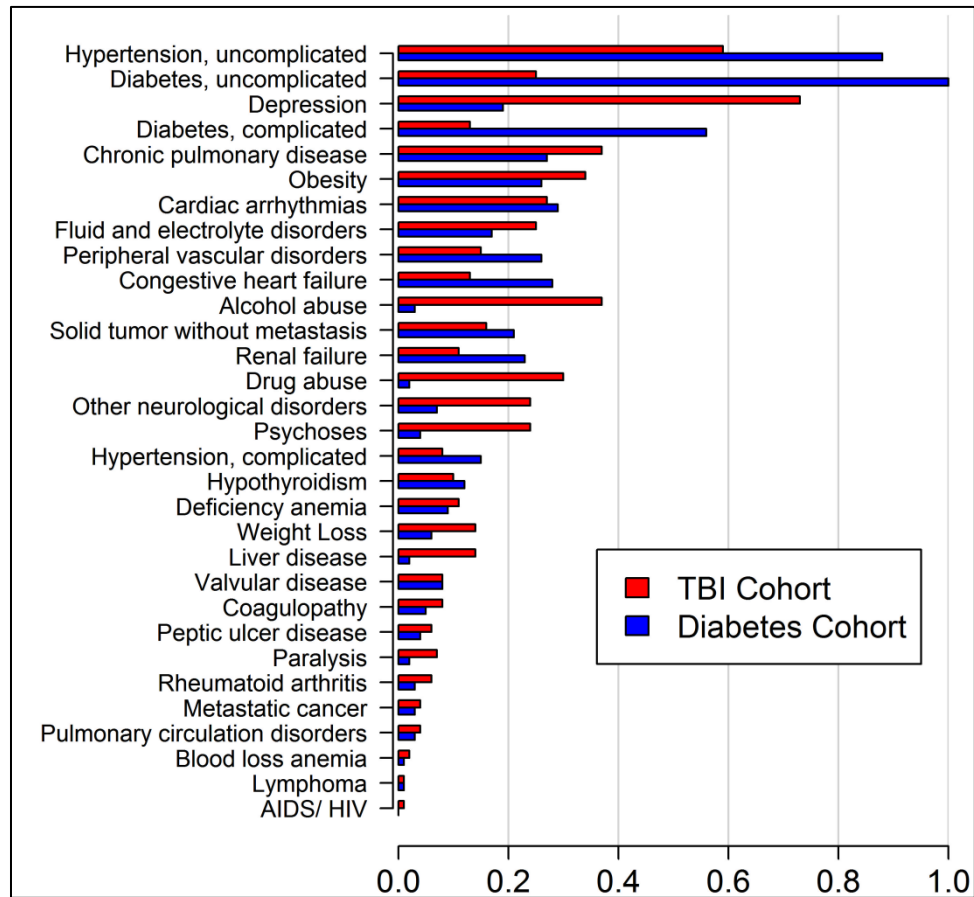


Figure 3: Prevalence of the 31 Elixhauser-Quan comorbidities in the diabetes and traumatic brain injury cohorts

models were less successful when compared to the Elixhauser-Quan model, particularly in the TBI population. In the TBI cohort, mean AUC values for unadjusted models varied between 0.83 and 0.92, compared with 0.83 for the Elixhauser-Quan model. In the DM cohort, they varied between 0.68 and 0.78, compared with 0.64 for Elixhauser-Quan model. In general, net reclassification statistics for predicting mortality (NRI (event)) for top-performing models were improved compared to the Elixhauser-Quan model, but NRI (non-event) results were either similar or slightly worse, indicating an improved ability to

Table 2: Mean performance statistics for Phase 1 models from validation data for the diabetes and traumatic brain injury cohorts based on 1000 replications. Unadjusted models were based only on ICD predictors, while adjusted models also included other patient demographic and clinical variables. The outcome was death within the study timeframe.

Diabetes Mellitus Cohort (Phase 1 models)									
		Elix-Quan	MARC	Assoc. Rules	Naïve Bayes	BART	Elastic-Net	Random Forest	Pooled
AUC	unadj.	0.64	0.68	0.71	0.72	0.75	0.75	0.77	0.78
	adj.	0.77	0.78	0.80	0.80	0.82	0.82	0.82	0.84
	adj.	0.73	0.74	0.76	0.75	0.77	0.78	0.78	0.79
NRI (event)	unadj.	ref	0.11	0.10	0.10	0.20	0.22	0.27	0.29
	adj.	ref	0.04	0.07	0.03	0.09	0.12	0.12	0.14
NRI (nonevent)	unadj.	ref	-0.05	-0.01	-0.02	-0.04	-0.05	-0.05	-0.05
	adj.	ref	-0.02	-0.01	0.03	0.00	-0.01	0.00	0.01
Brier Score	unadj.	0.23	0.22	0.21	0.21	0.20	0.20	0.19	0.19
	adj.	0.19	0.18	0.17	0.18	0.17	0.17	0.16	0.16
Traumatic Brain Injury Cohort (Phase 1 models)									
		Elix-Quan	MARC	Assoc. Rules	Naïve Bayes	BART	Elastic-Net	Random Forest	Pooled
AUC	unadj.	0.83	0.83	0.83	0.87	0.92	0.91	0.91	0.92
	adj.	0.87	0.89	0.88	0.88	0.92	0.92	0.91	0.93
NRI (event)	unadj.	ref	0.14	0.06	0.20	0.22	0.27	0.24	0.27
	adj.	ref	0.03	0.00	0.00	0.10	0.14	0.10	0.12
NRI (nonevent)	unadj.	ref	-0.03	-0.02	-0.10	0.01	0.00	0.00	0.00
	adj.	ref	0.00	-0.01	-0.02	0.01	0.00	0.01	0.01
Brier Score	unadj.	0.13	0.13	0.13	0.13	0.09	0.09	0.09	0.09
	adj.	0.11	0.11	0.11	0.11	0.09	0.09	0.09	0.08

predict which patients would die, but no improvement for predicting which patients would survive.

Table 3 and Figure 6 provide the results from phase 2, where the top-performing methods in phase 1 were applied (BART, elastic-net, RF, and pooled models). Here the

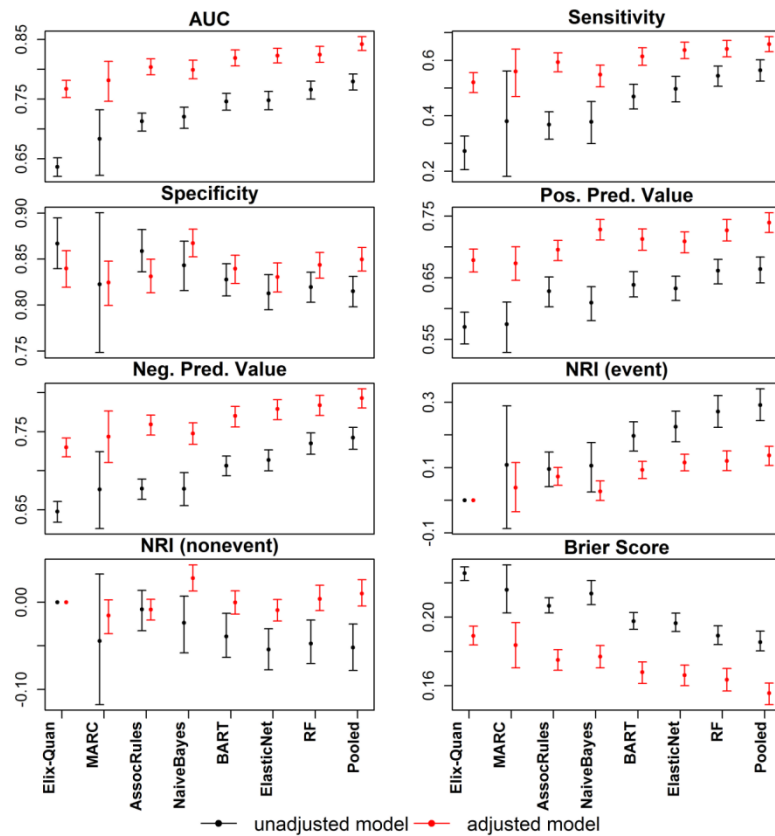


Figure 4: Diabetes Cohort performance statistics are shown for the unadjusted (in black) and adjusted (in red) phase 1 models, with 95% confidence intervals based on 1000 iterations. Adjusted models are based on both ICD code predictors and patient demographic variables, while unadjusted models are based only on ICD code predictors. The outcome was death within the study timeframe. NRI values for the Elix-Quan models are 0.00 since they serve as the reference.

outcome was five-year mortality rather than death within the study's timeframe, and ICD codes recorded within one year of death were excluded in order to avoid favoring conditions such as palliative care that would be strongly associated with death but would provide little long term predictive ability. As in phase 1, each model was trained in a single-disease dataset but was validated on a combined dataset comprised of the DM and TBI groups. This permitted a better evaluation of predictive performance in a more

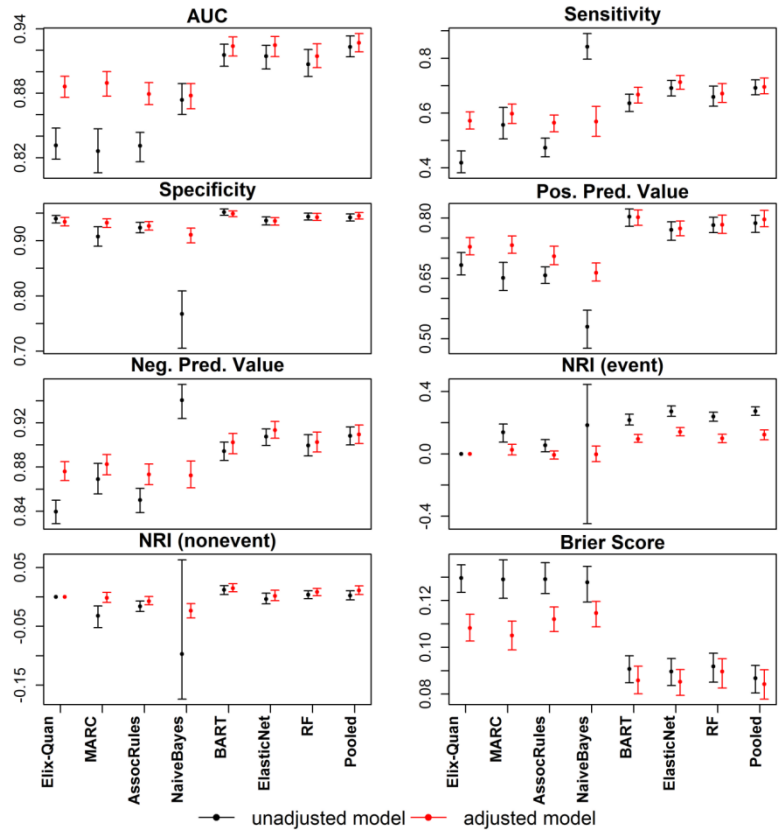


Figure 5: Traumatic Brain Injury Cohort: performance statistics are shown for the unadjusted (in black) and adjusted (in red) phase 1 models, with 95% confidence intervals based on 1000 iterations. Adjusted models are based on both ICD code predictors and patient demographic variables, while unadjusted models are based only on ICD code predictors. The outcome was death within the study timeframe. NRI values for the Elix-Quan models are 0.00 since they serve as the reference.

general population with a wider range of comorbidities. Phase 2 included only ICD codes as predictors since the phase 1 results demonstrated that adjusted for other covariates did not provide additional predictive performance insights. The BART, RF, elastic-net and pooled models were again superior to the Elixhauser model as seen in consistently higher mean AUC values and lower Brier scores for unadjusted and adjusted models in both cohorts. Similar to phase 1 results, the NRI statistics indicate performance gains

Table 3: Mean performance statistics from validation data for phase 2 models for diabetes and traumatic brain injury cohorts based on 1000 replications. Each model was trained on a single-disease dataset but was validated on a combined group drawn equally from the DM and TBI datasets. All models were unadjusted, based only on ICD predictors. The outcome was five-year mortality.

	Elixhauser- Quan	BART	Elastic- Net	Random Forest	Pooled
Diabetes cohort (training) with combined cohort (validation) (Phase 2)					
AUC	0.74	0.85	0.84	0.83	0.86
NRI (event)	Ref	0.08	0.10	0.10	0.16
NRI (nonevent)	Ref	-0.01	-0.01	-0.01	-0.01
Brier Score	0.07	0.07	0.07	0.07	0.06
TBI cohort (training) with combined cohort (validation) (Phase 2)					
AUC	0.74	0.89	0.88	0.83	0.89
NRI (event)	Ref	0.27	0.23	0.20	0.29
NRI (nonevent)	Ref	-0.02	-0.02	-0.01	-0.02
Brier Score	0.07	0.06	0.06	0.06	0.06

were seen in predicting mortality but little improvement was seen in predicting survival. In the TBI cohort, mean AUC values varied between 0.83 and 0.89, compared with 0.74 for the Elixhauser-Quan model. In the DM cohort, AUC values varied between 0.83 and 0.86, compared with 0.74 for Elixhauser-Quan model. Table 4 provides a summary of comorbidities that were important predictors of five-year mortality but which were not accounted for by the Elixhauser-Quan index. These comorbidities were identified by finding the common group of ICD-9-CM codes in the RF, elastic-net, and BART models for both the TBI and DM populations

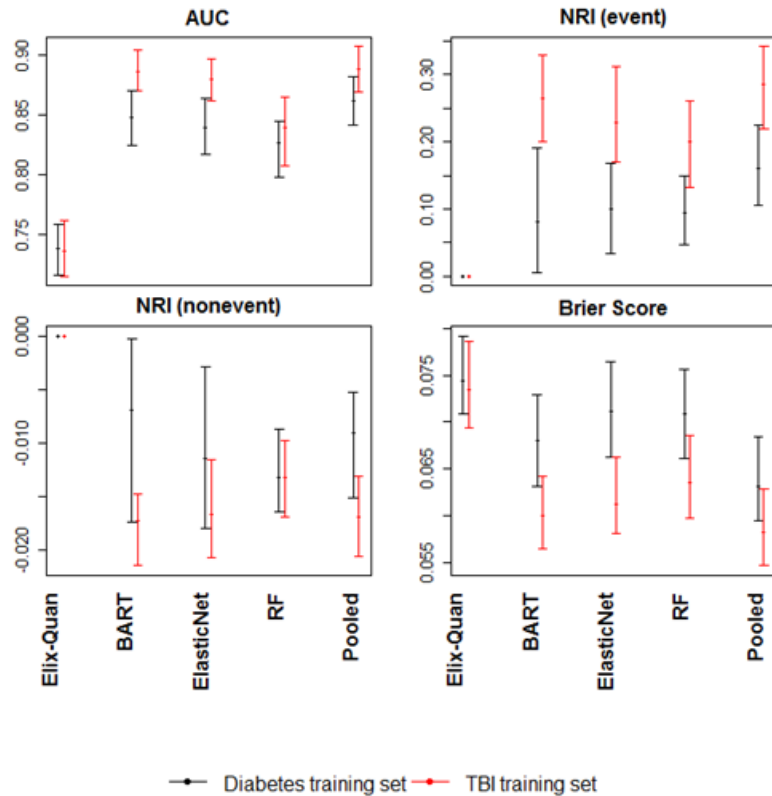


Figure 6: Phase 2 model mean performance statistics are shown with 95% confidence intervals based on 1000 iterations. Each model was trained on a single-disease dataset but was validated on a combined group drawn equally from the DM and TBI datasets. All models were unadjusted, based only on ICD predictors. The outcome was five-year mortality

which were not included in the Elixhauser index definition, were associated with mortality rather than survival, and were ranked in the top 50% for variable importance by each phase 2 model. Many of these conditions are related to functional status or cognitive problems.

Table 4: Summary of comorbidities from phase 2 models which were found to be important predictors of five-year mortality but were not accounted for by the Elixhauser-Quan index. These comorbidities were identified by finding the common group of ICD-9-CM codes associated with mortality in the RF, elastic-net, and BART models for both the TBI and DM populations that were consistently ranked in the top 50% for variable importance. This list was further narrowed to include only those conditions not accounted for by the Elixhauser-Quan index. Many of the below conditions are related to functional status or cognitive problems.

ICD-9-CM code	Description
07051	Acute hepatitis C
29048	Vascular dementia
2948	Other persistent mental disorders
3310	Alzheimer's disease
4111	Intermediate coronary disease
41400	Coronary atherosclerosis of unspecified type of vessel
4293	Cardiomegaly
436	Acute but ill-defined cerebrovascular disease
43889	Other late effects of cerebrovascular disease
5234	Chronic periodontitis
5251	Loss of teeth
5939	Disorder of kidney and ureter, unspecified
600	Hyperplasia of prostate
7070	Pressure ulcer
7809	Altered mental status
7866	Swelling, mass, or lump in chest
7872	Dysphagia
7993	Debility, unspecified
V048	Need for prophylactic vaccination against viral diseases
V604	No other household member able to render care
V651	Person consulting on behalf of another person
V670	Follow-up examination following surgery

3.4 Discussion

This work compared the performance of seven approaches for predicting patient outcomes based on comorbidities derived from ICD-9-CM codes. To the best of the author's knowledge, this is the first effort to conduct a detailed comparison of statistical and machine learning methods in the development of an improved prediction model based on the ICD system. In the first phase, the outcome was death within the study's timeframe, and no ICD-9 codes in the patient's record were excluded from consideration. Models were validated with observations for other patients drawn from the same cohort. The second phase involved a more robust evaluation of the top-performing models from the first phase. The outcome was five-year mortality, ICD-9 codes recorded within one year of death were excluded, and models were validated on a combined dataset drawn from both disease populations.

In both phases, the BART, RF, elastic-net and pooled models consistently had better predictive performance compared to models based on the Elixhauser-Quan index. Each method may have succeeded due to different strengths, of which none were seen in the Elixhauser-Quan approach. The pooled model, which attempted to merge the strengths from individual models, appeared to offer the best results in both phases and in both populations. This is consistent with conclusions that ensemble methods often outperform any single classifier [19]. The elastic-net model provided a balanced approach for handling possible collinearities between ICD-9 predictors while also shrinking less important estimated coefficients towards zero. The successful machine learning approaches (RF and BART) may automatically account for complex interactions that might otherwise be overlooked by other methods. Additionally, most models

attempted to account for a condition's relative severity rather than considering each to be an equally weighted and independent predictor.

Though the phase 2 models involved greater prediction challenges due to the shift to five-year mortality, the exclusion of ICD-9 data within a year of death, and the use of a wider population for validation, no substantial loss in predictive performance was seen. Mean AUCs for unadjusted DM models improved in phase 2 by 6% to 10%, while mean AUCs for unadjusted TBI models were 3% to 8% lower than for the corresponding phase 1 model. In all cases, each model had substantially better predictive performance when compared to the corresponding Elixhauser-Quan models. This provides some evidence these methods could be generalized to a wider population and to a range of different outcomes.

In addition to improved predictive performance, the phase 2 results provided additional insights into the patient populations beyond those provided by the Elixhauser index. As seen in Table 4, the patient's functional status was an important predictor of five-year mortality not accounted for by the Elixhauser-Quan model; examples of these conditions include cognitive problems, pressure ulcers, and caregiver status. This conclusion concerning functional status is consistent with previous research [61, 62]. Other serious conditions not included in the Elixhauser-Quan index were also identified; examples include Alzheimer's disease, cardiomegaly, and acute hepatitis C. These were likely excluded from the Elixhauser Index because they were not highly associated with the short-term outcomes used in its development.

Although the pooled model is based on a simple logistic regression, efforts to develop more complex ensemble models with improved prediction performance did not

lead to any improvements. For example, the prediction densities from the BART, RF and elastic net models were plotted separately for true positive, false positive, true negative and false negative training data observations under the pooled model. Differences in respective densities between the four groups were used to adjust the pooled model predictions, but this led to a slight drop in predictive performance. For example, any prediction improvement in the false negative group was negated by a decline in the true negative group.

There are several important limitations. First, this work was limited to two populations of generally older, male Veterans, and it was not demonstrated whether these methods would achieve similar results in other groups. Next, the use of administrative data imposes substantial risks for measurement inaccuracies and missing data. For example, one patient might have different ICD-9-CM codes entered for the same condition. In some cases less severe comorbidities such as diabetes, depression, angina or high blood pressure may be omitted from the record for critically ill patients; as a result, these conditions have been incorrectly associated with lower mortality odds in some studies [2]. Additionally, patients with good functional status and access to healthcare are more likely to have detailed health information recorded, while patients who are housebound, live in isolated rural areas, have cultural obstacles, or are otherwise disadvantaged are more likely to have incomplete records. Despite these sources of potential bias, a large body of previous work has shown that meaningful inference is possible from these data.

4. Second Manuscript: Improved Comorbidity Summary Score for Measuring Disease Burden and Predicting Outcomes with Applications to Three National Cohorts

4.1. Introduction

Research involving administrative healthcare data to study patient outcomes requires the investigator to carefully consider the patient's comorbidities, or disease burden in order to reduce the potential for biased inferences. This paper focuses on developing an improved summary score using one of the most popular sources for comorbidity information, that encoded by the International Classification of Diseases (ICD), which, in the ICD-9-CM version, consists of more than 14,000 unique codes. Each patient may have hundreds of ICD codes recorded over many years, and a large database may contain thousands of unique codes. Summary measures based on dimension reduction have thus become very popular tools. In some cases, these measures consist of a collection of disease conditions that serve as independent predictors; the Elixhauser comorbidity index is perhaps the most well-known example [2]. Other measures consist of single scores, such as the Charlson comorbidity index [1]. Because the Charlson and Elixhauser indices are well known and have been widely applied, investigators frequently use them without consideration for whether other methods could better adjust for disease burden.

ICD codes are primarily recorded for billing purposes, which can introduce numerous challenges when they are applied in research. Some disease conditions are

found to be under-reported when ICD codes are compared to the patient's clinical record [1, 63, 64]. Further, some chronic conditions such as high blood pressure or obesity are more likely to be recorded for patients who are generally fit, but more likely to be omitted for patients who are critically ill. Investigators have shown how this can lead to the false conclusion that some chronic conditions are associated with lower mortality odds [2]. This observation supported a hypothesis examined in this paper that gains in ICD summary measure performance might be made by including codes in prediction models that are associated with survival even when the clinical evidence suggests such codes may actually be associated with mortality.

The Charlson and Elixhauser indices were developed by different approaches, and were initially used to predict different outcomes. Charlson et al. [1] collected the comorbidities observed in 607 patients with hospital admissions during one month in 1984; these patients were then followed for one year. She used those baseline comorbidities to predict time to death over the one year period using Cox proportional hazards models. The relative magnitudes of the estimated coefficients were used to develop a weighted score that was validated in a population of 685 breast cancer patients. Elixhauser later noted that the Charlson score was soon repurposed by other investigators to predict numerous events other than one-year mortality, including short-term outcomes such as in-hospital mortality, hospital charges or length of stay. She was also concerned that the range of Charlson comorbidities was limited by the small population used to develop the score. Elixhauser et al. [2] instead considered the full range of conditions included in the ICD-9-CM coding manual as well as the comorbidities considered in a number of current studies. Her models were limited to predicting short-

term hospitalization outcomes, and used a narrow comorbidity definition that excluded conditions related to the primary reason for hospitalization, problems that might be complications that arose during treatment, or conditions she considered unimportant. To assess which conditions were most predictive, she conducted ordinary least squares regression or logistic regression to predict hospitalization charges, length of stay, or in-hospital mortality. She proposed an index of 31 independent comorbidities, and deliberately avoided combining them into a single score because she considered each investigator should examine the independent contributions of comorbidities where possible. Her index did not include numerous serious conditions that were not strongly associated with her short-term hospitalization outcomes; examples include dementia, Alzheimer's disease, or some types of renal disease. She reported that by excluding any condition that could be considered a complication of treatment, her index was less successful in predicting mortality than her other short-term outcomes. Such excluded conditions included pneumonia, cardiac arrest, cardiogenic shock, and respiratory failure [2].

The Elixhauser index has subsequently been applied to a wide range of outcomes, in some cases with little apparent regard for the reasoning behind the index's construction. For example, Baldwin et al. [65] used the Elixhauser index in models to predict two-year non-cancer mortality and the receipt of chemotherapy in cancer patients; Chu et al. [10] used it to predict one-year mortality; Lix et al. [66] used it to predict amputation, end stage renal disease, and stroke in diabetes patients. Since the Elixhauser index deliberately omitted numerous conditions not strongly associated with

short-term hospitalization outcomes or associated with treatment complications, the index may be less effective in predicting these other outcomes.

Both the Charlson and Elixhauser indices were developed using simple regression and proportional hazards models, and both involved arbitrary inclusion or exclusion of specific conditions rather than relying on strictly empirical methods to determine which predictors were most important. Since their development, numerous advances in statistical methods, machine learning algorithms, and computational power have occurred. In the first manuscript, a number of machine learning and statistical classification methods were compared to demonstrate that substantial improvements in prediction performance over existing indices could be achieved to predict five-year mortality. The most effective of these methods included Bayesian Additive Regression Trees (BART), Random Forest (RF), and elastic-net penalized generalized linear models. However, such models included up to 1000 predictors and could be cumbersome to apply in some areas of research. Here the goal is to instead develop a simple comorbidity summary score based on the insights gained in the previous work and show it has superior predictive performance to the Quan version of the Elixhauser index [9] when used to predict five-year mortality. The Elixhauser index is used as the basis for comparison because it was shown to have better predictive performance than the Charlson Index; the Quan version is used because of its improved performance over earlier versions of the Elixhauser Index [9]. The ICD-9-CM is included here rather than the ICD-10-CM system because the Veteran's Administration data used here was recorded under the older ICD-9 system, though the same methods could easily be applied to ICD-10-CM data.

4.2. Study Design and Methods

4.2.1 Study Populations

Three national cohorts of U.S. Veterans were used in this work; these had been created for earlier studies by linking numerous Veterans Health Administration patient and administrative databases. The first included 625,903 patients with diabetes mellitus (DM) based on two or more related ICD-9-CM codes and at least one prescription filled for a medication to treat diabetes [47]. In the original study, Veterans were followed from 2002 until death, loss to follow-up, or until December 2006, and newer data was added here to extend the follow-time until December 2012. The second cohort involved 168,125 Veterans diagnosed with traumatic brain injury (TBI) during 2004 and 2005. In the original study, patients were followed from the point of entry until death, loss to follow-up, or until December 2010 [48]; newer data was added to extend the follow-time to December 2012. The third cohort involved 3,359,560 patients with chronic kidney disease (CKD) defined for stages 1 through 5 based on estimated glomerular filtration rates calculated from serum creatinine levels and the patient's age, gender, and race. CKD patients were also identified through ICD-9-CM codes. CKD patients were followed from 2000 until December 2012, loss to follow-up, or until death. Kidney or liver transplant recipients were excluded (M. N. Ozieh, M. Gebregziabher, R. Ward, D. J. Taber, L. Egede, unpublished data, 2016). All studies were approved by the Medical University of South Carolina Institutional Review Board (IRB) and the Ralph H. Johnson Veterans Affairs Medical Center Research and Development committee.

4.2.2 Covariates and outcome

The predictors in all models were binary variables, each based on a single ICD condition. No other predictors were included since it was demonstrated in the first manuscript that no additional insights into effective modeling of ICD data were obtained by including other patient covariates. The outcome for all models was five-year mortality.

4.2.3 Methods

Figure 7 provides an overview of the index development steps. For each patient in the three disease cohorts, the available ICD-9-CM codes were collected from the earliest available data until one year prior to mortality or through the end of the study if the patient did not die. Codes recorded within a year of death were excluded because conditions that often occur in this period (such as palliative care) might provide an unrealistic advantage against the Elixhauser models but would provide little help in making long-term predictions. An early cutoff for starting ICD collection was not established; instead all available codes were used for each patient. The challenge for the competing models was similar to asking, “given all of the patient’s ICD codes up until today, predict whether he or she will die in the next five years.” Although patients varied by the length of available ICD code history, this was considered this was an expected condition in chronic disease cohorts, and results were compared to those from Elixhauser models that faced the same challenges.

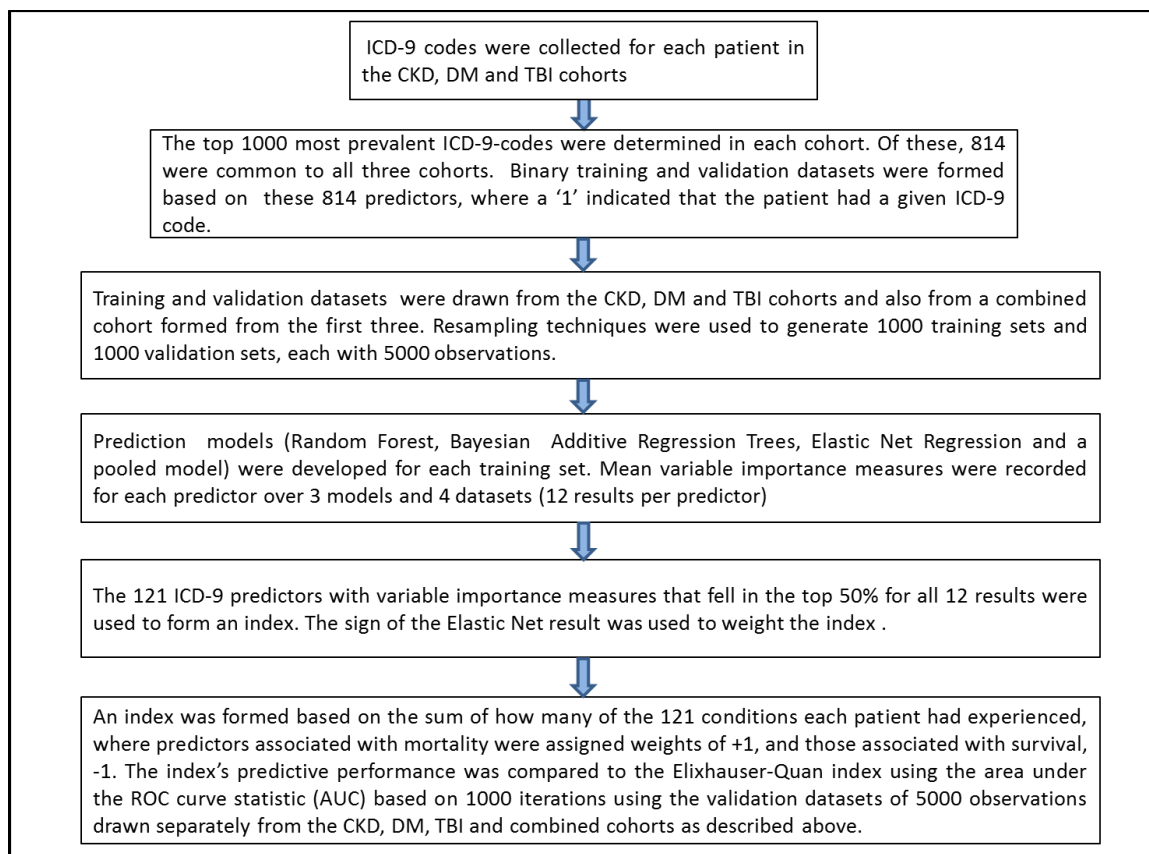


Figure 7: Overview of summary score development.

Prior to analysis, erroneous duplicate formats of ICD-9 codes were identified and corrected to the single correct format. For example, codes 250, 2500, and 25000 all represent the same condition, but each would be treated as separate predictors in machine learning algorithms. Out of the approximately 14,000 unique ICD-9 codes, the 1000 most prevalent codes for each cohort were retained, and then those codes that were not common to all cohorts were excluded. There were 814 such codes, which formed the common set of binary predictors, where a '1' indicated the given ICD-9-CM code was found in a patient's record.

Four types of training and validation data sets were generated; three were drawn from the DM, CKD and TBI cohorts, and the fourth was a combined cohort created by sampling in approximately equal proportions from the three disease cohorts. Use of the combined cohort was intended to show how well the index might function in a more general population.

Resampling methods were used due to concerns for computational efficiency with the machine learning methods, for which it could be difficult or impossible to complete an analysis of the entire dataset without use of a parallel computing environment. Smaller test and training datasets were generated, each with 5000 observations, by randomly sampling the full datasets 1000 times with replacement. Performance statistics were collected for each validation run and the overall mean and 95% confidence intervals generated by 1000 iterations were used to compare the models' relative performance. As demonstrated by Marshall et al. [37] and Gebregziabher et al. [49], this non-parametric bootstrapping approach is reasonable for our large datasets, such that independence between numerous samples is reasonably assured.

4.2.3.1 Prediction Models

The top-performing methods from the first manuscript: random forest (RF), Bayesian additive regression trees (BART) and elastic-net penalized regression (REG), are again used here to provide variable importance measures for use in summary score development. The machine learning methods (RF and BART) are capable of automatically accounting for complex interactions between predictors that are likely to

exist in these data. Elastic-net penalized regression provides an efficient way to handle possible collinearities between predictors, while shrinking the estimated coefficients of less important predictors. Each method is used to develop separate estimates of variable importance for use in index development.

Random forest [28] is an ensemble method based on classification trees that is often effective in datasets with many weak predictors, as is the case with ICD-9 data. It relies on bootstrap aggregation (or ‘bagging’) to generate a forest, which is termed “random” due to the random selection of a pre-specified number of features (or predictor variables) at each tree’s nodes. The feature that leads to the largest improvement in the tree’s classification ability is then used to split the data at that node. The random forest method can automatically account for complex interactions, and was reported to be very competitive with other machine learning methods when compared on the basis of misclassification error [14]. Each variable’s mean decrease in the Gini Index is used as a measure of variable importance for use in model development. For N_m observations at node m , for outcome variable y with class levels k , and with predictors $x \in R_m$, the proportion of observations at a node for a given predictor and class level k is

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) .$$

The Gini Index at this node is given by [14] :

$$G = \sum_{k=1}^K \hat{p}_{mk} (1 - \hat{p}_{mk}) .$$

Large changes in the Gini index at a given node indicate the splitting variable’s importance is relatively high, and the mean decrease over all nodes involving a given

predictor across all trees and across 1000 iterations is used as the variable importance measure. These values are always positive, regardless of whether a predictor is associated with survival or mortality. The R package *randomForest* is used to implement this method [58].

Bayesian Additive Regression Trees (BART) [60] is an extension of the supervised tree-based ensemble learning method, but unlike random forest, prior distributions are established for each tree's decision rules and terminal node parameters, and an MCMC algorithm is used to sample from the posterior distribution for the ensemble of trees. The authors contend their approach provides a substantial degree of regularization such that each tree's complexity is reduced. Predictive results reported in some datasets were superior to random forest, neural nets, and regularized regression methods [60].

The R package *BayesTree* was used to implement BART [59]. This algorithm develops the model based on the training data, and then provides the results of post-convergence samples from the posterior distribution using test data. Variable importance is estimated by the mean count of how many times each predictor is selected for use in a node's decision rule among all trees over all MCMC samples and over all 1000 iterations. These values are always positive.

Elastic-net regression, which involves the use of a regularized generalized linear model, provided another measure of variable importance. In the work supporting the first manuscript a number of regularized regression methods were compared, including ridge regression [50], LASSO regression [51], elastic-net regression [52], and group LASSO

regression [53], and found that elastic-net regression provided the best predictive performance in the ICD code data. The elastic-net model combines the LASSO and ridge approaches with the addition of parameter α such that the loss function becomes the LASSO model for $\alpha = 1$, and the ridge model when $\alpha = 0$. For binary outcome $y \in (-1, 1)$, predictors $x_i : (1, x_{1i}, \dots, x_{p-1,i})$ and shrinkage parameter λ , the following equation is minimized in order to determine coefficient estimates:

$$- \min \left[\sum_{i=1}^n \log(1 + e^{-y_i \sum_{k=1}^p \beta_k x_{ik}}) + \lambda \sum_{j=1}^p \{ \alpha |\beta_j| + (1 - \alpha) \beta_j^2 \} \right].$$

$\alpha = 0.5$ provided the best predictive performance. The R package *glmnet* [54] was used to implement this method, and used the mean coefficient estimates from 1000 iterations as variable importance measures. The sign of the mean coefficient estimates were used to weight the index by ± 1 based on the association with survival or mortality, as discussed further below.

4.2.3.2 Score algorithm:

When analyses were completed on the DM, TBI, CKD and combined datasets, there were 12 variable importance results from the RF, BART and REG models for each of the 814 ICD-9-CM predictors. The following algorithm was used to develop a summary score:

- (1) Determine which predictors have variable importance measures ranked in the top 50% in all 12 results.

(2) For p such predictors and the i^{th} patient, establish $d_{ij} : d_{i1}, d_{i2}, \dots, d_{ip}$,
 $d \in (0,1)$, to signify which predictors are recorded in a given patient's record.

(2) Determine whether each predictor is associated with mortality or survival based on whether the majority of elastic-net estimated parameters from the four datasets are positive or negative; in the rare case of a tie assume a mortality association. Note that RF and BART variable importance measures are always positive regardless of the association.

(3) Assign weights $w : \{w_1, w_2, \dots, w_p\}$ of +1 to those conditions associated with mortality and -1 to those associated with survival.

(4) Calculate a summary score for each patient: $S_i = \sum_{j=1}^p w_j d_{ij}$.

4.2.4.3 Score assessment

The summary score was used as the single predictor in a logistic regression models using validation datasets from each of the four population groups. Its performance was compared to similar models based on the Elixhauser index using the area under the ROC curve (AUC), net reclassification improvement, Brier score [22], sensitivity and specificity statistics.

Table 5: Demographic information for the Chronic Kidney Disease, Diabetes and Traumatic Brain Injury cohorts.

Variable	Level	Chronic Kidney Disease	Diabetes	Traumatic Brain Injury
Mean age		75.0	73.1	49.9
Five-year mortality (%)		41.9	38.7	20.9
Gender (%)	male	96.7	97.0	93.7
	female	3.3	3.0	6.3
Marital status (%)	single	6.3	7.0	26.0
	widowed	14.8	11.0	4.6
	divorced	21.0	21.0	25.9
	married	58.0	59.0	42.1
Race/ethnicity (%)	Non-Hispanic white	81.4	76.0	55.8
	Non-Hispanic black	13.6	15.0	13.0
	Hispanic	2.9	5.0	1.9
	other or missing	2.2	4.0	29.2
Homeless (%)		6.3	8.0	1.5
Greater than 50% disability (%)	service-related	23.4	27.0	23.3

4.3. Results

Table 5 provides a summary of demographic information for the three cohorts. Five-year mortality ranged between 20.9 and 41.9%. The TBI cohort's mean age was 49.9, while the other groups had mean ages of 75.0 and 73.1. The groups' gender and racial-ethnic makeup is typical for Veteran populations with these age distributions. Between 23% and 27% of Veterans had at least 50% disability connected with their military service.

Figure 8 and Table 6 compare the performance of the summary score to the Elixhauser index based on area under ROC curve (AUC), Brier score, net

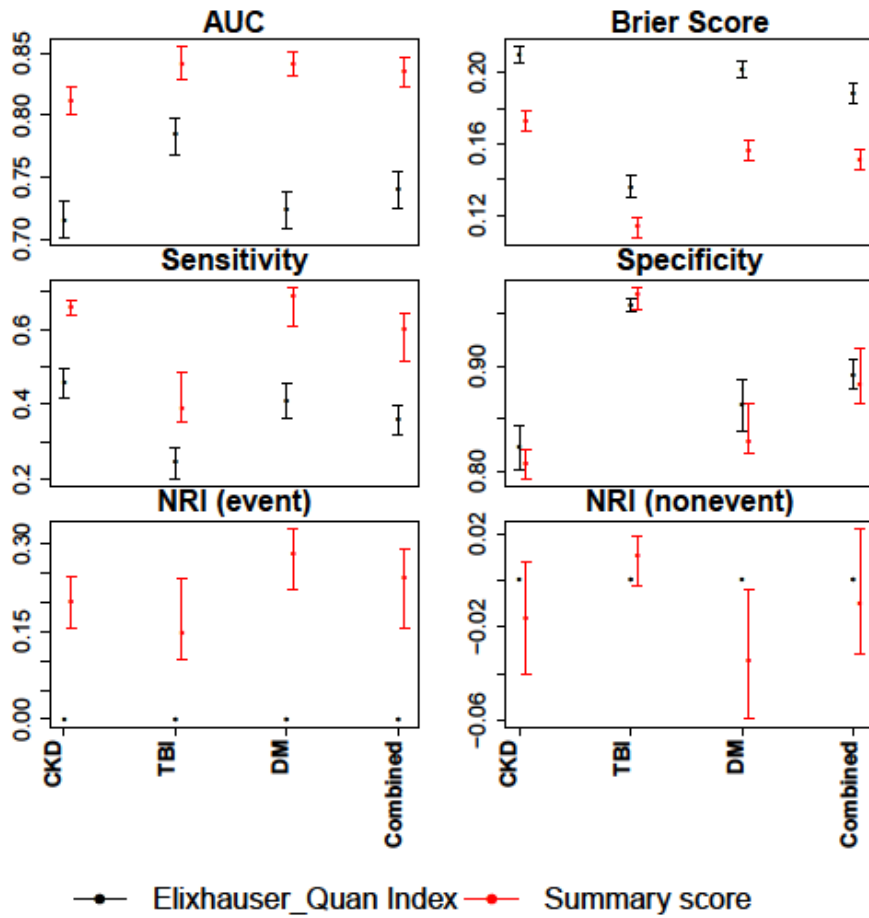


Figure 8: Mean performance statistics with 95% confidence intervals based on 1000 iterations using validation datasets for the chronic kidney disease (CKD), traumatic brain injury (TBI), type 2 diabetes mellitus (DM), and the combined cohort, which was formed by randomly drawing from the first three cohorts in equal proportions. AUC is the area under the receiver operator characteristic curve; NRI (event) and NRI (nonevent) are the net reclassification improvement statistics for mortality and survival. NRI values for the Elixhauser models are set at 0.00 since they serve as the reference. The outcome was five-year mortality. In each cohort, summary score models demonstrated significantly better predictive performance compared to models based on the 31 Elixhauser comorbidities.

reclassification improvement, sensitivity, and specification statistics from models validated on each disease cohort and on a combined cohort. Mean AUC values for the four datasets ranged between 0.81-0.84 for the new score, and between 0.72 - 0.78 for

Table 6: Mean performance statistics with 95% confidence intervals based on 1000 iterations using validation datasets for the chronic kidney disease (CKD), traumatic brain injury (TBI), type 2 diabetes mellitus (DM), and the combined cohort, which was formed by randomly drawing from the first three cohorts in equal proportions. AUC is the area under the receiver operator characteristic curve; NRI (event) and NRI (nonevent) are the net reclassification improvement statistics for mortality and survival. NRI values for the Elixhauser models are set at 0.00 since they serve as the reference. The outcome was five-year mortality. In each cohort, summary score models demonstrated significantly better predictive performance compared to models based on the 31 Elixhauser comorbidities.

Cohort	CKD		TBI	
Index	Elix-Quan	Summary Score	Elix-Quan	Summary Score
AUC	0.72 (0.70; 0.73)	0.81 (0.80; 0.82)	0.78 (0.77; 0.80)	0.84 (0.83; 0.85)
Sensitivity	0.46 (0.42; 0.50)	0.66 (0.64; 0.68)	0.24 (0.20; 0.29)	0.39 (0.35; 0.49)
Specificity	0.82 (0.80; 0.84)	0.81 (0.79; 0.82)	0.96 (0.95; 0.96)	0.97 (0.95; 0.97)
NR(event)	ref	0.20 (0.16; 0.24)	ref	0.15 (0.10; 0.24)
NRI(nonevent)	ref	-0.02 (-0.04; 0.01)	ref	0.01 (0.00; 0.02)
Brier Score	0.21 (0.20; 0.21)	0.17 (0.17; 0.18)	0.14 (0.13; 0.14)	0.11 (0.11; 0.12)
Cohort	DM		Combined	
Index	Elix-Quan	Summary Score	Elix-Quan	Summary Score
AUC	0.72 (0.71; 0.74)	0.84 (0.83; 0.85)	0.74 (0.73; 0.75)	0.84 (0.82; 0.85)
Sensitivity	0.41 (0.36; 0.45)	0.69 (0.61; 0.71)	0.36 (0.32; 0.40)	0.60 (0.51; 0.64)
Specificity	0.86 (0.84; 0.89)	0.83 (0.82; 0.86)	0.89 (0.88; 0.91)	0.88 (0.86; 0.92)
NR(event)	ref	0.28 (0.22; 0.33)	ref	0.24 (0.15; 0.29)
NRI(nonevent)	ref	-0.03 (-0.06; 0.00)	ref	-0.01 (-0.03; 0.02)
Brier Score	0.20 (0.20; 0.21)	0.16 (0.15; 0.16)	0.19 (0.18; 0.19)	0.15 (0.15; 0.16)

the Elixhauser index. Brier score values were consistent with the AUC results. Mean sensitivity values for the new score ranged between 0.39 – 0.69, compared to 0.24 – 0.46 for the Elixhauser index. Mean specificity values for the score ranged between 0.81 – 0.97, compared to 0.82 – 0.96 for the Elixhauser models. Mean net reclassification improvement (NRI) statistics for predicting mortality or survival were consistent with the respective trends in sensitivity and specificity.

Table 7 (following the Discussion section) lists the ICD-9-CM codes used to predict five-year mortality in the summary score, and indicates whether each code is included in the Elixhauser index's definition. Overall, 97 of the 121 codes in the index were not part of the Elixhauser index definition, and 63 of the 121 codes are associated with mortality.

4.4. Discussion

This work produced a comorbidity summary score for predicting five-year mortality that had stronger predictive performance than the widely-used Elixhauser index. The models used in the score's development were trained and validated on three large Veterans Administration datasets and further validation was based on a combined cohort, which provided a broad range of comorbidities and disease severity levels. The score was comprised of ICD-9 codes with variable importance measures that fell in the top 50% of all twelve model runs (four training datasets and three classification methods). Strong improvements in predictive performance were demonstrated based on AUC and Brier Score statistics. There were also some improvements in sensitivity and net reclassification improvement for mortality when compared to the Elixhauser index, while specificity values remained generally the same. The score's strong performance in the combined cohort provided some initial evidence that it could be successfully applied to a more general population, but further work is needed to demonstrate this.

The summary score approach differs from existing measures in several ways:

(1) First, 58 of 121 conditions were negatively weighted since they were associated with survival in the models, and many patients had an overall negative score. As hypothesized, by including codes statistically associated with survival, model predictive performance improved, even when the clinical evidence suggests such codes in some cases may actually be associated with mortality. Such codes may be recorded in healthier patients who are not being treated for more serious conditions (obesity or hyperlipidemia for example); other codes in this category simply recorded routine outpatient visits (routine screening or exam codes). The models predicted that patients with large numbers of these negatively weighted conditions and few of the more serious illnesses are more likely to survive.

(2) Next, the summary score is simpler to implement since it consists of only 121 ICD-9 codes, compared to more than 1000 unique codes in the Elixhauser index definition.

(3) While the Elixhauser index definition excluded conditions not associated with short-term hospitalization outcomes and any acute conditions considered treatment complications, all such conditions that occurred at least one year prior to death were considered since they could be valid mortality predictors. This approach is more suitable for the long-term outcome was considered here. Examples of acute conditions included in the summary score but excluded in the Elixhauser index include pneumonia and acute cerebrovascular disease.

(4) The summary score contained a number of conditions related to the patient's functional status that were not covered by the Elixhauser index, including Alzheimer's

disease, senile dementia, hearing loss, persistent mental disorders, memory loss, falls, no other household member able to render care, and urinary incontinence.

Several conditions in the summary score may have unexpected associations with mortality, and warrant further discussion. For example, nail dermatophytosis, or nail fungal infection, might generally be considered a benign condition, but Scher et al. [67] and Loo [68] report its prevalence rises both with age and the presence of peripheral vascular disease and diabetes. This may explain its predictive importance in these Veteran populations. In another example, the code 'V048: need for prophylactic vaccination against other viral diseases' might also be considered to be benign, but further investigation showed this code may be a proxy for age since its use was discontinued in 2003 when it was replaced by a number of other codes [69].

The new comorbidity measure is not intended to provide a comprehensive clinical summary of a patient's disease burden; instead, it provides a simple prediction of five-year mortality based on a comparison with millions of other Veterans for whether a the patient has specific conditions that were most predictive in this population.

Although summary scores are convenient tools, investigators should apply them with care. As Elixhauser et al. noted [2], combining individual predictors into a single index may lead to a loss of explanatory power. Romano et al. [8] commented that summary indices might be most appropriate in small datasets where it is not feasible to model a large group of comorbidities. They also warned that investigators should not apply an index without carefully considering the assumptions and outcomes used in its development; this is a warning that appears to be unheeded by many investigators.

Scneeweiss et al. [70] further cautioned that the weights developed for one population are not likely to be generalizable to other groups. Although the new score was developed using three large datasets involving different disease cohorts with a wide variety of comorbidities, these populations are generally limited to older male Veterans, and further work is needed to determine the score's predictive performance in a wider population. As Scneeweiss et al. [70] wrote, a summary score might be most suitable for use as a convenient data exploration tool to rapidly assess large ICD code datasets. In general, investigators working with such data may be most successful by developing dedicated comorbidity models for their unique populations and outcomes using the methods described here.

Table 7: ICD-9-CM conditions that form the summary score, ordered by ICD hierarchy. Those conditions contained in the Elixhauser index definition [9] are indicated by a “+” symbol. Conditions associated with mortality have an index weight of +1; those associated with survival have weights of -1. 97 of the 121 codes in the index were not part of the Elixhauser index definition, and 63 of the 121 codes are associated with mortality. Many of those conditions not included in the Elixhauser index definition are related to the patient’s functional status.

Contained in Elixhauser Definition	Index Weight	ICD-9-CM code	Condition
	1	1101	Dermatophytosis of nail
+	1	1629	Malignant neoplasm of bronchus and lung, unspecified
	1	1733	Unspecified malignant neoplasm of skin, unspecified parts of face
	1	1739	Other malignant neoplasm of skin
+	1	185	Malignant neoplasm of prostate
+	1	25000	Diabetes mellitus without mention of complication
+	1	25001	Type I Diabetes mellitus
+	1	25060	Diabetes with neurological manifestations
	-1	2722	Mixed hyperlipidemia
	-1	2724	Other and unspecified hyperlipidemia
+	1	2765	Volume depletion disorder
+	1	2767	Hyperpotassemia
+	-1	27800	Obesity, unspecified
+	1	2809	Iron deficiency anemia, unspecified
+	1	2859	Anemia, unspecified
+	1	2875	Thrombocytopenia, unspecified
	1	2900	Senile dementia, uncomplicated
	1	2948	Other persistent mental disorders
	1	2949	Unspecified persistent mental disorders
+	1	2989	Unspecified psychosis
	-1	30272	Psychosexual dysfunction
	-1	32723	Obstructive sleep apnea
	1	3310	Alzheimer's disease
+	1	3320	Parkinson’s disease
	-1	33829	Other chronic pain
	-1	3540	Carpal tunnel syndrome
	1	36201	Background diabetic retinopathy
	1	36250	Macular degeneration (senile), unspecified
	1	36251	Nonexudative senile macular degeneration
	-1	36501	Open angle glaucoma with borderline findings, low risk
	1	36610	Senile cataract, unspecified

Table 7 (continued)

Contained in Elixhauser Definition	Index Weight	ICD-9-CM code	Condition
	-1	3671	Myopia
	-1	38830	Tinnitus, unspecified
	-1	38831	Subjective tinnitus
	1	38910	Sensorineural hearing loss, unspecified
	1	3892	Mixed conductive and sensorineural hearing loss
+	1	40391	Hypertensive chronic kidney disease with end stage renal disease
	1	41400	Coronary atherosclerosis
+	1	4241	Aortic valve disorders
+	1	42731	Atrial fibrillation
+	1	4280	Congestive heart failure, unspecified
	1	436	Acute, but ill-defined, cerebrovascular disease
	1	4389	Unspecified late effects of cerebrovascular disease
+	1	4439	Peripheral vascular disease, unspecified
	1	4538	Acute venous embolism and thrombosis of other specified veins
	1	45981	Venous (peripheral) insufficiency, unspecified
	-1	4619	Acute sinusitis, unspecified
	-1	462	Acute pharyngitis
	-1	4659	Acute upper respiratory infections of unspecified site
	-1	4739	Unspecified sinusitis (chronic)
	-1	4779	Allergic rhinitis, cause unspecified
	1	486	Pneumonia, organism unspecified
+	1	49121	Obstructive chronic bronchitis with (acute) exacerbation
+	1	4928	Other emphysema
+	-1	49390	Asthma, unspecified type, unspecified
+	1	496	Chronic airway obstruction
	1	51889	Other diseases of lung, not elsewhere classified
	-1	52102	Dental caries extending into dentine
	-1	52103	Dental caries extending into pulp
+	1	5715	Cirrhosis of liver without mention of alcohol
	1	5789	Hemorrhage of gastrointestinal tract, unspecified
+	1	585	Chronic kidney disease
	-1	5920	Calculus of kidney
	1	5939	Unspecified disorder of kidney and ureter
	1	5990	Urinary tract infection
	1	5997	Hematuria
	1	7051	Acute hepatitis C without mention of hepatic coma
	-1	71536	Osteoarthritis, localized, lower leg
	-1	71941	Pain in joint, shoulder region
	-1	71944	Pain in joint, hand
	-1	71946	Pain in joint, lower leg
	-1	71947	Pain in joint, ankle and foot

Table 7 (continued)

Contained in Elixhauser Definition	Index Weight	ICD-9-CM code	Condition
	-1	7231	Cervicalgia
	-1	7242	Lumbago
	-1	7243	Sciatica
	-1	72690	Enthesopathy of unspecified site
	-1	72871	Plantar fascial fibromatosis
	1	73300	Osteoporosis, unspecified
	-1	78057	Unspecified sleep apnea
	1	78097	Altered mental status
	1	7812	Abnormality of gait
	-1	7820	Disturbance of skin sensation
	1	7823	Edema
	1	78321	Loss of weight
	-1	7840	Headache
	-1	78659	Other chest pain
	1	78820	Retention of urine, unspecified
	1	78830	Urinary incontinence, unspecified
	-1	79021	Impaired fasting glucose
	-1	79029	Other abnormal glucose
	1	7931	Abnormal findings on radiological /other examination of lung field
	-1	7962	Elevated blood pressure reading without diagnosis of hypertension
	1	7993	Debility, unspecified
	-1	9953	Allergy, unspecified, not elsewhere classified
	1	E8889	Unspecified fall
	-1	V0382	Other vaccinations against streptococcus pneumoniae
	1	V048	Need for prophylactic vaccination, other viral diseases
	-1	V0481	Need for prophylactic vaccination and inoculation against influenza
	-1	V065	Need for prophylactic vaccination against tetanus-diphtheria
	-1	V1272	Personal history of colonic polyps
	1	V431	Lens replaced by other means
	-1	V531	Fitting and adjustment of spectacles and contact lenses
	1	V583	Attention to dressings and sutures
	1	V5861	Long-term (current) use of anticoagulants
	-1	V5883	Encounter for therapeutic drug monitoring
	1	V604	No other household member able to render care
	-1	V653	Dietary surveillance and counseling
	-1	V6540	Counseling NOS
	-1	V6549	Other specified counseling
	-1	V659	Unspecified reason for consultation

Table 7 (continued)

Contained in Elixhauser Definition	Index Weight	ICD-9-CM code	Condition
	-1	V6801	Disability examination
	1	V681	Issue of repeat prescriptions
	-1	V700	Routine general medical examination at a health care facility
	-1	V703	Other general medical examination for administrative purposes
	-1	V705	Health examination of defined subpopulations
	-1	V7189	Observation and evaluation for other specified suspected conditions
	-1	V7260	Laboratory examination, unspecified
	-1	V7651	Special screening for malignant neoplasms of colon
	-1	V802	Screening for other eye conditions
	-1	V812	Screening for other and unspecified cardiovascular conditions
	-1	V8289	Special screening for other specified conditions

5. Third Manuscript: Comprehensive Comparison of Machine Learning and Model-Based Multiple Imputation Methods with Competing Sensitivity Analyses for Non-Random Missingness.

5.1. Introduction

Missing data is a frequent problem in administrative healthcare databases, and investigators working with such data must carefully assess how to best approach this problem in order to reduce the possibility for biased results. In Veterans Health Administration (VHA) research to investigate the reasons for health inequities among minority groups, missing data in key variables such as patient race and ethnicity can pose tremendous challenges. In past years, researchers often dealt with the missing data problem by simply conducting complete-case analysis, though this strategy could lead to biased results unless the data were missing completely at random. More recently, steps to attempt to assess the pattern of missingness and methods to help achieve unbiased results such as multiple imputation are commonly seen.

When assessing missing data, it is important to determine what type of relationship exists between the missing values and the mechanism that led to their being missing. Three such scenarios are typically defined [71]:

- a. Missing completely at random (MCAR): in this situation, the probability of missing values does not depend on either the observed or the missing values.
- b. Missing at random (MAR): In this case, the probability of missing values depends on the observed values, but does not depend on the missing values.

- c. Missing not at random (MNAR): in this case, the probability of missing values occurring depends on unobserved observations. The MNAR pattern cannot be ruled out by examining the data since it exists due to information not contained in the data, and investigators who rely on imputation methods that rely on MAR or MCAR assumptions should take additional steps to assess whether their results are sensitive to changes under MNAR conditions.

Numerous parametric imputation methods exist for handling data with MCAR or MAR patterns; multiple imputation by chained equations (MICE) is one commonly used approach due to its ability to handle multiple imputation for mixed data types [25, 26]. MICE imputes missing values from separate distributions for each variable with missing values conditional on the other variables, but has been criticized for lacking a theoretical basis [27], and for requiring the investigator to have advance knowledge of non-linear relationships or collinearities between predictors [17]. Other researchers have concluded that machine learning methods can automatically handle interactions and other concerns while also producing inference estimates with narrower confidence limits and with more computational efficiency. The random forest algorithm has been applied in several multiple imputation research efforts, and involves bootstrap aggregation of numerous independent decision trees, and can account for complex interactions and collinearities between predictors more readily than many parametric methods, while the ensemble voting of independent trees naturally lends itself to an efficient imputation process [28]. For example Stekhoven et al. [16] claim their multiple imputation approach (missForest) based on the random forest method was superior to traditional statistical methods including MICE, based on improved misclassification error rates or normalized root

mean squared errors. Jerez et al. [29] provided a similar conclusion based on a comparison of machine learning and statistical imputation methods. Other researchers have incorporated machine learning methods within an existing statistical method; for example, Shah et al [17] incorporated random forest as the multiple imputation method within the existing MICE method and showed the new approach had a superior ability to handle nonlinear relationships and collinearities.

Though the multiple imputation methods described above are capable of producing unbiased results under MCAR and MAR, such results are far less likely when a missing not at random (MNAR) condition exists. As Verbeke et al. [30] discuss, it is possible to construct models based on MNAR assumptions, but these assumptions are not testable since their support is not contained in the data. Further, Molenburghs et al. [31] demonstrated that it is not possible to empirically distinguish between MNAR and MAR situations from the data alone because for every MNAR model, it is possible to build an MAR model with the same fit. The most common approach given these circumstances is to conduct sensitivity analysis on MAR models to examine their stability when MNAR assumptions are introduced [32, 33]. Though numerous approaches are possible, two general types of sensitivity analyses are most common; these are based on pattern mixture models [32 – 34] and selection models [35].

5.1.1 Motivating Example

This research involves a VHA cohort of 161,586 Veterans treated for traumatic brain injury (TBI) between 2004 and 2010. In the original study, patients were followed from the point of entry until death, loss to follow-up, or until December 2010 [48]; newer

data was merged to extend the follow-time to December 2012. The study was approved by the Medical University of South Carolina Institutional Review Board (IRB) and the Ralph H. Johnson Veterans Affairs Medical Center Research and Development committee.

The original dataset had approximately 30% missing race-ethnicity, which was derived solely from VHA Corporate Data Warehouse (CDW) MedSAS files. The missing proportion was reduced to 2% by merging newer information from the CDW PatSub_PatientRace and Pat_Sub_PatientEthnicity tables, followed by Medicare race-ethnicity information from the VitalStatus table. Table 8 provides the demographic and clinical characteristics for this group, and Table 9 provides a comparison of the original and updated race distributions. The newer race distribution appears to be more typical of the VA population, and the combined effects of better race-ethnicity data collection [24] and the use of Medicare data [23] provide solid support for the claim that the newer distribution is more accurate. The updated race distribution provided strong evidence that an MNAR pattern existed in the original data. It was then possible to compare the results of several multiple imputation methods using the original data against results obtained by using the updated race distribution. Since the motivating example involves MNAR missingness, this research also involved applying several types of sensitivity analyses to determine if such approaches provide any additional insights.

Consistent with the problems seen in the TBI cohort, several investigators have reported the absence of race information in VHA or Medicare data may be due to non-random causes [4 – 6]. Depending on the timeframe being studied, the level of missingness may be substantial in VHA data. Stroupe et al. [23] reported that 48% of

VHA patient records had missing race-ethnicity information in 2004, but this value had been reduced to 15% by 2012 [3] due to concerted efforts to collect this information and due to a 2003 requirement for recording self-reported race-ethnicity rather than observer-reported values [24]. Stroupe et al. [23] demonstrated that further improvements were possible by merging VHA data with Medicare data; in the author's experience with several VHA cohorts followed through 2012 or later, the missing race fraction can be reduced below 3% in some cases.

This research provides a unique contribution by conducting a comprehensive comparison of multiple imputation (MI) methods under both MAR and MNAR conditions using approaches that incorporate both machine learning and statistical methods. Additionally, for missing race/ethnicity variables under MNAR, it examines the effectiveness of several types of sensitivity analyses, both in simulations and in real data application. The remainder of this aim is organized as follows: the Methods section provides a description of the multiple imputation methods and sensitivity analyses that are applied here, first in a simulation, and then in the TBI example. The simulation framework is then described, including how MCAR, MAR and MNAR patterns are generated in the simulation data. In the Results and Discussion sections the insights gained from this work are reviewed, particularly that MNAR missingness is an extremely challenging problem, even when the data's MNAR mechanism is well understood.

5.2. Methods

Multiple imputation (MI) is a common approach for handling missing data. Rubin et al. [25] provided a detailed description of multiple imputation's advantages over

numerous single imputation methods, particularly that MI is far more capable of modeling the uncertainty associated with each imputed value, especially when the reason for value being missing is unknown. Van Buuren et al. [26] discussed the challenges for applying MI to multivariate data, where many predictors can have missing values, and described two general approaches:

- a. Joint modeling, where a joint parametric multivariate distribution is specified, and imputations are generated in a Bayesian framework from the posterior predictive distribution. However, this approach requires the analyst to fully specify the model, such that any unknown interaction or nonlinearity may lead to biased results.
- b. Fully conditional specification models (FCS), or Multiple Imputation by Chained Equations (MICE): here each predictor has a distribution conditional on all of the other predictors, with distribution parameters specific to each predictor rather than associated with a joint distribution. This provides the important advantage of being able to easily handle continuous and categorical data types since each predictor has its own conditional distribution [26]. In MICE models, if \mathbf{Y} is a matrix for n patients and k predictors, and a portion of each predictor is missing:

\mathbf{y}_j^{obs} are the observed observations for the j th predictor,

\mathbf{y}_j^{mis} are the missing observations for the j th predictor, and

\mathbf{y}_{-j} is defined as all of the predictors except the j th predictor.

For $P(\mathbf{Y}_j | \mathbf{Y}_{-j}, \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is the vector of parameters for the j th conditional distribution, each of the k parameters and predictors is successively sampled via a Gibbs sampler, where the t^{th} iteration is represented by:

$$\theta_j^{*(t)} \sim P(\theta_j | Y_j^{obs}, Y_{-j}^{(t-1)})$$

$$Y_j^{*(t)} \sim P(Y_j^{mis} | Y_j^{obs}, Y_{-j}^{(t-1)}, \theta_j^{*(t)})$$

M sampling processes are conducted in parallel, where M is typically 5 or 10, and each is continued for enough iterations to ensure convergence, typically less than 20 iterations. M imputed data sets are produced, and the results are pooled as follows [25] :

$$\bar{\theta}_j = \sum_{i=1}^M \frac{\hat{\theta}_i}{M}, \text{ where } \bar{\theta}_j \text{ is the pooled parameter estimate.}$$

$$\bar{U}_j = \sum_{i=1}^M \frac{U_i}{M} \text{ where } \bar{U}_j \text{ is the arithmetic average of the parameter estimate}$$

variances.

$$B_j = \sum_{i=1}^M \frac{(\hat{\theta}_i - \bar{\theta}_j)^2}{M-1} \text{ where } B_j \text{ is the variance of the parameter estimates.}$$

$$T_j = \bar{U}_j + (1 + M^{-1}) * B_j \text{ where } T_j \text{ is the total pooled variance for the } j^{th} \text{ pooled parameter.}$$

5.2.1 MICE methods

Numerous imputation methods can be applied within the MICE framework, including traditional statistical methods and machine learning algorithms. The *MICE* package in R [72] was used to implement this framework, and separate functions were written to incorporate the BART and neural net methods. The R program code developed here is made available for downloading as described in Appendix A.

5.2.1.1 MICE with logistic regression and predictive mean matching (MICE-LR)

Here, logistic regression and predictive mean matching [73] are used to form the predictive conditional distributions for missing categorical and continuous variables. The goal was to compare these traditional methods with the machine learning approaches described below.

5.2.1.2 MICE with random forest (MICE-RF)

Here the random forest algorithm [28] is used during each imputation to generate a selected number of trees based on observed data, each of which is used to make a prediction. The imputed value is randomly selected from these predictions.

5.2.1.3 MICE with Bayesian Additive Regression Trees (MICE-BART)

BART [59] is an extension of the supervised tree-based ensemble learning method, but unlike random forest, prior distributions are established for each tree's decision rules and terminal node parameters, and an MCMC algorithm is used to sample from the posterior distribution for the ensemble of trees. The R packages *BayesTree* [60] and *MPBART* [74] were used to generate imputed values for continuous and categorical variables, respectively. A separate function was developed here to incorporate these methods within the MICE framework. While this method produced reasonable results for several iterations of the simulation, it was too slow to be viable when used on a typical PC, though it could be useful in a parallel environment.

5.2.1.4 MICE with neural net (MICE-NNET)

Here, a single hidden layer neural network was used during each imputation step [14]. This machine learning approach is a non-linear statistical model that approximates more traditional classification and regression models. The R package *nnet* [75] was used to implement this method, and the size and weight decay parameters were tuned using 10-fold cross validation with the *caret* package in R [76]. A separate function was developed to incorporate neural net imputations within the MICE framework.

5.2.2 Random forest multiple imputation:

The random forest algorithm was applied as a multiple imputation method independent of the MICE framework using the *missForest* package in R [16]. This algorithm orders the predictors with missing values based on their increasing proportion of missing values, and then imputes each variable in turn by generating a random forest based on the observed values for the variable of interest and all corresponding observations from the other variables. This forest is then used to impute the missing values of the variable of interest. Once each variable has been imputed, the entire process is repeated until a stopping point is reached based on the difference between successive imputed datasets.

5.2.3 MNAR sensitivity analyses

Several sensitivity analyses were compared; these involve imposing MNAR assumptions on multiple imputation models that are based on MAR assumptions. This was of particular interest since there was strong evidence that the race-ethnicity

covariate in the TBI example had an MNAR missingness mechanism. Each of these methods requires the investigator to make assumptions about the missingness pattern, though such assumptions are unverifiable from the data itself. The departure of MNAR sensitivity analysis results from the MAR results is an indication that the MAR assumption may not hold.

5.2.3.1 Pattern mixture model adjustment [32]

A pattern mixture model assumes that a number of missingness patterns may exist, each with a separate joint distribution for the partially and fully observed variables. For patients $i = 1, \dots, n$ and covariates Y_{1i} and Y_{2i} , assume Y_{1i} has missing values with indicator R_i , such that $R_i = 0$ when Y_{1i} is missing and $R_i = 1$ otherwise. Under MNAR the joint distribution $f(Y_{1i}, Y_{2i}, R_i)$ is factored as $f(Y_{1i}, Y_{2i} / R_i) f(R_i)$, where the joint distribution of the partially and fully observed variables is conditional on the partially observed variable.

Since the MNAR distribution cannot be determined from the observed data, Carpenter and Kenward [32] suggest starting from the MAR scenario and then adjusting the model using MNAR assumptions in order to examine whether the MAR model is sensitive to such changes. For example, the race-ethnicity variable in the TBI data has 4 levels, and a multinomial logistic model was used to impute the missing values under MAR assumptions, where the probability for imputing race group level j is given by:

$$pr(\text{race} = j) = \frac{e^{d_j}}{\sum_{k=1}^4 e^{d_k}}, \text{ where}$$

$d_k = \alpha_k + \mathbf{x}'\boldsymbol{\beta}_k$, for $k < 4$ and $d_4 = 0$, and where α and β are multinomial model parameters.

In order to test various MNAR assumptions, shift parameters δ_k are introduced for each level of the race variable, and the probability for imputing race group level j becomes:

$$pr(\text{race} = j) = \frac{e^{d_j + \delta_j}}{\sum_{k=1}^4 e^{d_k + \delta_k}} .$$

Following the adjustment, we then examine how the model inference changes under the MNAR assumption. An iterative processes is used to determine the combination of shift parameters that best matches the MNAR assumptions.

Several other types of pattern mixture models are applied to MNAR sensitivity analysis. One group of such methods involves data with monotone missingness patterns, which are defined for variables y_1, \dots, y_p , such that when y_j is missing for a given observation, then it is also missing for y_k with $k > j$. Such observations are grouped based on their missingness patterns, and specific groups are then used to impute a given variable. Two such methods were attempted here; the first is termed complete case missing values (CCMV), in which only observations with no missing values are used [73]. The second is neighboring case missing value (NCMV), where, for imputing values of y_j , the closest group in the monotone hierarchy is used for imputation [77]. In this closest group, observations exist for y_j but not for y_{j+1} .

5.2.3.2 Parameter re-weighting

In this selection model approach described by Carpenter et al. [36], the estimated parameters determined from multiple imputation datasets generated under MAR assumptions are reweighted to reflect MNAR assumptions. This approximation requires that the MAR and MNAR distributions for these parameters overlap.

The MNAR assumption is incorporated in a logistic model,

$$\text{logit}(\text{Pr}(R_i = 1)) = \alpha + \beta'X_i + \delta Y_i,$$

where the outcome that patient i has an observed value for covariate Y is related to δY_i such that large positive values of δ make the odds for observing Y in this patient much higher, while large negative values have the opposite effect. Carpenter et al. [36] show that for m imputations and $i = 1, \dots, n_i$ patients who are missing covariate Y , the weight for the m^{th} imputation is related to a linear combination of the imputed data:

$$\tilde{w}_m = \exp\left(\sum_{i=1}^{n_i} -\delta Y_i^m\right), \text{ and the normalized weight is } w_m = \frac{\tilde{w}_m}{\sum_{i=1}^m \tilde{w}_m}. \text{ The imputed results are}$$

pooled in a manner similar to that developed by Rubin [25]:

$$\hat{\beta}_{MNAR} = \sum_{m=1}^M w_m \hat{\beta}_m, \text{ where } \hat{\beta}_m \text{ is the MAR parameter estimate for the } m^{\text{th}} \text{ imputation;}$$

$$\bar{U} = \sum_{m=1}^M w_m \hat{\sigma}_m^2 \text{ where } \bar{U} \text{ is the weighted mean of parameter estimate variances;}$$

$$B = \sum_{m=1}^M w_m (\hat{\beta}_m - \hat{\beta}_{MNAR})^2 \text{ where } B \text{ is the between variance of the parameter estimates.}$$

$$T = \bar{U} + (1 + M^{-1}) * B \text{ where } T \text{ is the total pooled variance.}$$

While Carpenter et al. demonstrated this approach for continuous outcomes with MNAR missingness, Heraud-Bousquet et al. [78] provide additional insights for applying the weighting method to datasets with missing covariates, including categorical variables.

5.2.1 Simulation Study

A simulation study was conducted in order to compare the multiple imputation methods described above against results from complete-case analysis under MCAR, MAR and MNAR scenarios. 1000 datasets with 5000 observations each were selected by randomly sampling with replacement from a Veterans Administration dataset that consisted of approximately 37,000 complete case observations from a diabetes cohort. Though it would have been possible to fully simulate such data, a resampling approach was used instead to help ensure that the complex structures and associations found in real patient observations were also present in the synthetic datasets. As demonstrated by Marshall et al. [37] and Gebregziabher et al. [38] this approach is reasonable when the original dataset is large enough to help assure independence between samples.

The outcome was mortality within the 10-year study timeframe, and covariates included demographic measures such as age (continuous variable), gender, racial-ethnic group (non-Hispanic white, non-Hispanic black, Hispanic, other), marital status (married or single), and urban-rural location indicator. Clinical indicators included the percentage of disability connected to military service, the patient's mean medication possession ratio (mean MPR), and the patient's mean glycated hemoglobin (mean A1c) level during the study period. The variables with missingness imposed were racial-ethnic group, mean A1c, and mean MPR.

MCAR, MAR and MNAR missingness scenarios were separately imposed on each bootstrapped dataset, and for each of these in turn, versions were generated with 10%, 30%, or 50% missing values. Complete case analysis was conducted on each of these nine datasets, along with multiple imputation by the four methods discussed above. Missing data patterns were generated by the following rules [79]:

- 1) Missing Completely at Random (MCAR): missing observations for the racial-ethnic group, mean MPR, and mean A1c variables were determined on a completely random basis.
- 2) Missing at Random by rank (MAR): when the patient died, the racial-ethnic group value was more likely to be missing; when the patient was single, the mean MPR variable was more likely to be missing; when the patient lived in a rural location, the mean A1c variable was more likely to be missing.
- 3) Missing Not At Random (MNAR): when the patient was in the non-Hispanic black or Hispanic groups and died during the study window, the racial-ethnic value for that patient was more likely to be missing. When mean MPR or mean A1c were above their respective medians, each was more likely to be missing.

Once missing values were established using the rules described above, further adjustments were made on a random basis as needed to achieve the required total proportion of observations with any missing values. Finally, each dataset was tested using logistic regression to verify that the required missingness structure had been generated. Binary indicators were generated for each of the variables with missing values, such that a '0' meant the value was missing. These indicators served as the

outcomes in the three logistic regression models. For each type of missingness, the significance of the estimated parameters was evaluated, and a given dataset was accepted if odds ratios for the parameters of interest were at least 1.5. For example, in the MNAR case, predictors of interest were the non-Hispanic black and Hispanic groups, and mortality.

Imputation methods were compared using the following statistics:

- 1) Relative bias: $(\hat{\beta} - \hat{\beta}^o) / \hat{\beta}$, where $\hat{\beta}$ and $\hat{\beta}^o$ are the generalized linear model parameter estimates based on the imputed data and the full dataset of 37,506 complete cases, respectively.
- 2) Efficiency: $\text{var}(\hat{\beta}) / \text{var}(\hat{\beta}^o)$
- 3) Root mean square error: $\sqrt{(\hat{\beta} - \hat{\beta}^o)^2 + \hat{\sigma}^2}$, where $\hat{\sigma}^2$ is the estimated variance of the parameter estimate from the model based on imputed data.
- 4) Coverage probability: the probability based on 1000 bootstrapped iterations that the 95% confidence interval for the parameter estimate contains $\hat{\beta}^o$.

5.3 Results

5.3.1 Simulation results:

Table 8 provides a summary of clinical and demographic characteristics for the diabetes (simulation) cohort. Figures 9 and 10 provide relative bias results for MAR and MNAR scenarios for those variables on which missingness was imposed: non-Hispanic black and Hispanic groups and for mean medication possession ratio (mean MPR) and mean glycated hemoglobin (mean A1c). Figure 11 provides coverage probability results.

When confidence intervals are compared in the MAR scenario, MICE with random forest imputation appeared to provide the least biased results when compared to complete case analysis, particularly at 50% missingness.

In the MNAR scenario, for the non-Hispanic black and Hispanic groups, all multiple imputation results were biased. However, for the two continuous variables, the MICE methods provided reasonable results. Coverage probability under MNAR is very poor for both race groups regardless of the MI method, but for mean MPR and mean A1C, coverage probability remains high. The missForest MI method appears to lag the other MI methods in coverage performance.

Figure 12 provides simulation results for MNAR sensitivity analyses, which were performed on multiple imputation results from data with 30% MNAR missingness in the non-Hispanic black and Hispanic groups. Under pattern mixture model 1 and selection model parameter weighting, shift parameters were iteratively adjusted to achieve the lowest relative bias when compared against the true race distribution. Under pattern

Table 8: Demographic and clinical characteristics of the diabetes (used in simulation) and the traumatic brain injury cohorts.

Variable	Level	Diabetes Cohort n = 37,506	TBI Cohort n = 161,586
Mean age (sd)	---	73.4 (5.4)	49.9 (17.9)
Mortality rate (%)	---	45.6	23.9
Gender (%)	male	98.7	93.7
Marital status (%)	non-married ²	28.2	57.9
	married	71.8	42.1
Race/ethnicity (%) ¹	non-Hispanic white	80.9	55.9
	non-Hispanic black	9.9	13.0
	Hispanic	5.1	1.9
	other	4.0	2.6
	missing	---	26.6
Rural location (%)	---	39.0	---
More than 50% service-related disability (%)	---	7.3	23.3
TBI severity	less severe	---	22.7
	moderate	---	27.5
	highest	---	49.8
Mean HbA1c (mean/sd) ³	---	7.2 (1.1)	---
Mean MPR (mean/sd) ⁴	---	0.79 (0.2)	---

¹ original race-ethnicity distribution in TBI cohort

² includes single, divorced, widowed, never married

³ mean glycosylated hemoglobin

⁴ mean medication possession ratio (number of days of diabetes medication supply divided by 365 days (or if deceased during that year, the number of days until death) over the study period

mixture model 2 (PMM-2), shift parameters were iteratively adjusted in order to provide the closest match in the imputed data with the true race distribution proportions for each group. While PMM-1 results did achieve low relative bias, the imputed datasets had substantially more non-Hispanic black and Hispanic members than seen in the original data. In PMM-2 on the other hand, when imputed datasets had approximately the same race distribution as in the original data, relative bias remained high. Bias was also high after parameter reweighting analysis.

Table 9: comparison of original and updated race-ethnicity distributions for the traumatic brain injury (TBI) data. NHW is non-Hispanic white, NHB is non-Hispanic black. Under the original distribution, 26% of patients were classified as missing race-ethnicity information; under the newer distribution, which incorporated more recent VHA data sources and merged Medicare information, the percentage of missing values was reduced to approximately 2%. The new race distribution showed that race was likely missing under MNAR conditions in the original data (see table 10).

Updated Distribution	Original distribution					Total	(percent)
	NHW	NHB	Hispanic	Other	Missing		
NHW	83190	335	621	0	29562	113708	0.70
NHB	478	19911	42	0	5795	26226	0.16
Hispanic	5565	385	2339	0	4267	12556	0.08
Other	1024	444	117	4231	0	5816	0.04
Missing	0	0	0	0	3280	3280	0.02
Total	90257	21075	3119	4231	42904	161586	
(percent)	0.56	0.13	0.02	0.03	0.27		

5.4.2 Results from TBI application:

The original and updated race and ethnicity distributions for the TBI group are shown in Table 9. By merging updated VHA and Medicare information, the proportion of missing values was reduced from about 26% to 2%. Of note, 34% of the Hispanic group in the updated distribution was in the missing category under the original distribution, compared with 26% and 22% for non-Hispanic whites and non-Hispanic blacks, respectively. Further, 47% of the Hispanic group had been misclassified to a different group originally, substantially higher than for other groups.

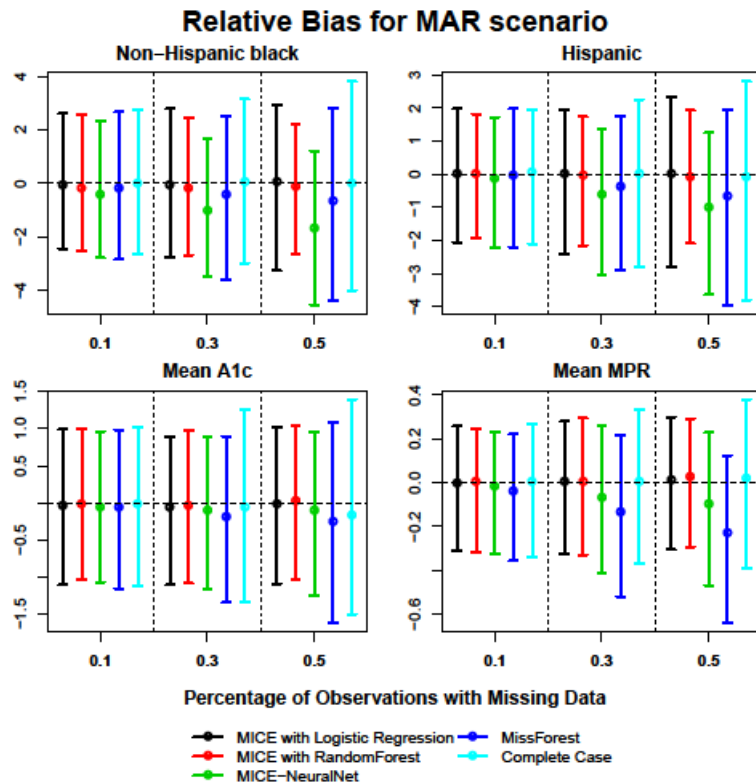


Figure 9: Simulation results: relative bias with 95% confidence intervals for four multiple imputation methods compared to complete case analysis under MAR missingness. In this scenario, when the patient died, the racial-ethnic variable was more likely to be missing; when the patient was single, mean medication possession ratio (mean MPR) was more likely to be missing; when the patient lived in a rural location, the mean glycated hemoglobin (mean A1c) variable was more likely to be missing. Based on relative confidence intervals, MICE with random forest imputation appeared to provide the least biased results when compared to complete case analysis, particularly for at 50% missingness.

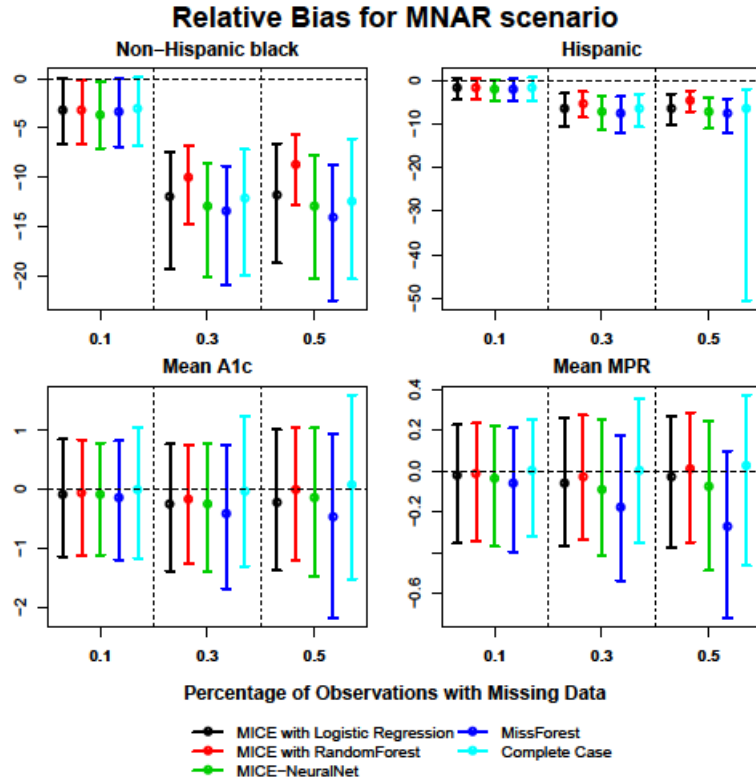


Figure 10: Simulation results: relative bias with 95% confidence intervals for four multiple imputation methods compared to complete case analysis under MNAR missingness. In this scenario, when the patient was in the non-Hispanic black or Hispanic groups and died during the study window, the racial-ethnic variable for that patient was more likely to be missing. When mean medication possession ratio (mean MPR) or mean glycated hemoglobin (mean A1c) were above their respective medians, each was more likely to be missing. For the non-Hispanic black and Hispanic groups, all of the multiple imputation results were biased. For the two continuous variables, however, the MICE methods provided reasonable results.

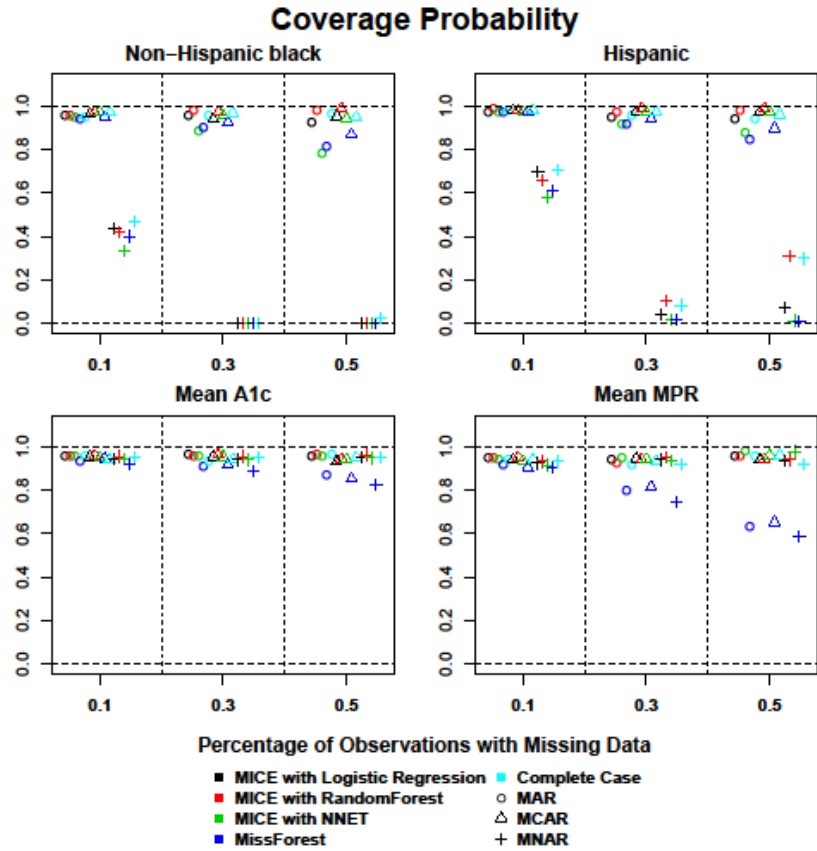


Figure 11: Simulation results: coverage probability for each multiple imputation (MI) method and missingness scenario. For the non-Hispanic black and Hispanic groups, coverage probability under MNAR is very poor regardless of the MI method. For the continuous variables, coverage probability remains high under MNAR. The missForest MI method appears to lag the other MI methods in coverage performance.

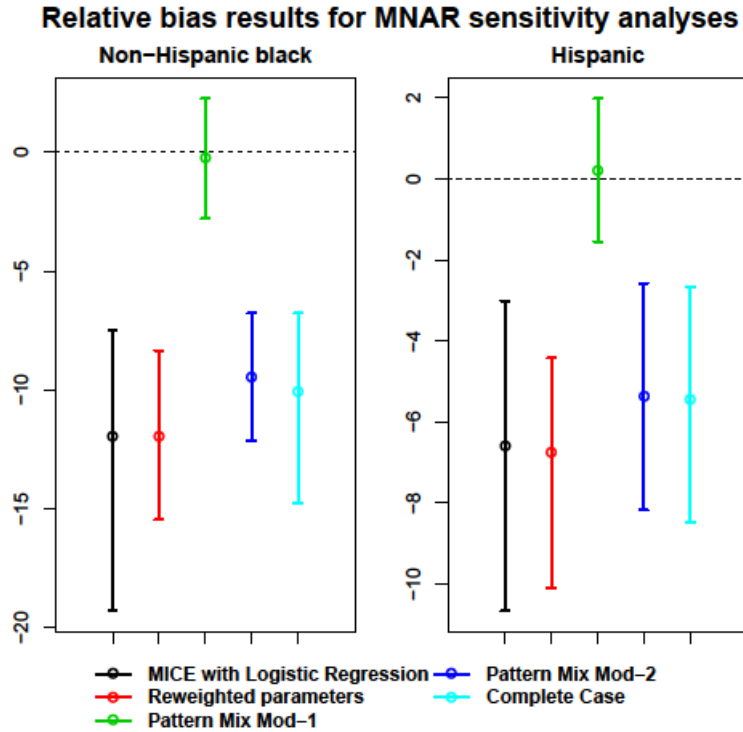


Figure 12: Simulation results for MNAR sensitivity adjustment: comparison of relative bias for three MNAR sensitivity analyses, all with 30% missingness. Sensitivity adjustments were made to multiple imputation results based on multinomial logistic regression models. These are compared with the unadjusted results (MAR imputation, in black) and with complete case analysis (light blue). Under pattern mixture model 1 and selection model parameter weighting, shift parameters were iteratively adjusted to achieve the lowest relative bias when compared against the true race distribution. Under pattern mixture model 2 (PMM-2), shift parameters were iteratively adjusted in order to provide the closest match in the imputed data with the true race distribution proportions for each group. While PMM-1 results were substantially improved, the imputed datasets had substantially more non-Hispanic black and Hispanic members than seen in the original data. In PMM-2, when imputed datasets had approximately the same race distribution as in the original data, relative bias remained high. The parameter reweighting analysis did not succeed because the distribution of MNAR coefficients fell outside the distribution of the MAR coefficients for any plausible adjustment.

Table 10 summarizes the association between missing race in the TBI data and other covariates. Of note, the association between the updated race-ethnicity variable and missing race in the original data shows indications of an MNAR pattern in the Hispanic group, with OR = 1.47 (95% CI, 1.41 - 1.52).

Table 11 compares the mortality odds ratios based on the updated race-ethnicity variable with those for complete case analysis in the original data, multiple imputation results, and three MNAR sensitivity analyses. The complete case odds ratio for Hispanics is 1.22 (95% CI, 1.11 - 1.33) compared with non-Hispanic whites, while the OR based on the updated race information is protective: 0.73 (95% CI, 0.69 - 0.77). In the multiple imputation comparison, the MissForest results appear to be biased lower for all three race groups, while the other multiple imputation methods provided generally similar results, and none differed substantially from complete case analysis. In the first pattern mixture model analysis (PMM-1) and selection model parameter weighting, shift parameters were iteratively adjusted to achieve the lowest relative bias when compared against the updated race distribution. The PMM-1 Hispanic OR result was 0.85 (95% CI: 0.82 – 0.88), but the imputed data contained an average of 28% Hispanic patients, compared with the actual value of 8%. The parameter weighting Hispanic result was nearly identical to the complete case result. Under PMM-2, shift parameters were iteratively adjusted in order to provide the closest match in the imputed data with the updated race distribution; here the Hispanic group OR was 1.04 (95% CI, 0.96 - 1.12), lower than the complete case result.

Table 10: Odds ratios for the association between missing race-ethnicity in the original TBI dataset and other covariates. The race-ethnicity predictor here is the updated, or “true” race determined with newer information. The strong association between the Hispanic group and missing race is strong evidence for an MNAR mechanism.

Variable	Level	OR (95% CI)
Age		1.00 (1.00, 1.00)
Gender	Female	1.08 (1.03; 1.13)
Marital Status	Married	---
	Non-married	0.88 (0.86; 0.91)
Race-ethnicity ¹	NHW	---
	NHB	0.91 (0.78; 0.83)
	Hispanic	1.47 (1.41; 1.52)
	Other ²	0
TBI severity	less	---
	moderate	1.27 (1.23; 1.31)
	most	1.34 (1.30; 1.39)
Homeless		0.41 (0.36; 0.47)
Death		0.78 (0.75; 0.80)
Disability >50%		0.82 (0.79; 0.85)

¹Using updated race-ethnicity distribution to predict missing values in older race-ethnicity data

²No patients were missing in the ‘other’ category

5.4 Discussion

Three MICE methods and the missForest algorithm were compared against complete case analysis in MCAR, MAR and MNAR scenarios and several types of MNAR sensitivity analysis were then applied, both in a simulation and in an application to TBI data. One specific goal was to examine competing methods for approaching the problem of missing race-ethnicity information typically seen in VHA datasets. In

Table 11: Traumatic brain injury (TBI) example: comparison of mortality odds ratio based on the updated race distribution against the odds ratios from complete case analysis, multiple imputation (MI), and MI with MNAR sensitivity analyses using the original TBI data, where approximately 26% of the race-ethnicity data was missing. Under PMM-1 and selection model parameter weighting (“weighted”), shift parameters were iteratively adjusted to achieve the lowest relative bias when compared against the updated race distribution. Under pattern mixture model 2 (PMM-2) shift parameters were iteratively adjusted in order to provide the closest match in the imputed data with the updated (“true”) race distribution.

Analysis	MI Type	MNAR sensitivity analysis	Mortality OR by race-ethnicity group (95% CI) Ref group is NHW		
			NHB	Hispanic	Other
OR based on updated race distribution	---	---	0.80 (0.77; 0.83)	0.73 (0.69; 0.77)	0.92 (0.85; 1.00)
Complete Case Analysis	---	---	0.79 (0.75; 0.83)	1.22 (1.11; 1.33)	0.71 (0.64; 0.79)
Multiple imputation	Multinomial logistic regression	---	0.76 (0.73; 0.80)	1.19 (1.10; 1.30)	0.71 (0.65; 0.78)
	Random Forest	---	0.84 (0.72; 0.97)	1.18 (0.86; 1.51)	0.78 (0.55; 1.08)
	Neural Net	---	0.85 (0.83; 0.88)	1.13 (1.02; 1.27)	0.71 (0.64; 0.81)
	MissForest	---	0.62 (0.60; 0.65)	0.93 (0.89; 0.98)	0.35 (0.32; 0.38)
MI with MNAR sensitivity analysis	Multinomial logistic regression	PMM-1	0.82 (0.79; 0.85)	0.85 (0.82; 0.88)	0.67 (0.60; 0.74)
		PMM-2	0.80 (.076, 0.77)	1.04 (0.96; 1.12)	0.68 (0.61; 0.75)
		Weighted	0.79 (0.76; 0.83)	1.21 (1.11; 1.31)	0.67 (0.60; 0.75)

particular, this work examined whether MI methods that incorporate machine learning algorithms have any performance advantage, and whether any sensitivity analyses were more successful under MNAR. The TBI example provided a good opportunity for this comparison.

As expected, under MCAR and MAR conditions in the simulation, the MICE methods and miss Forest imputation provided reliable results with reasonably low bias and good efficiency when compared with complete case analysis. In particular, MICE with random forest imputation appeared to have a slight performance advantage, while the missForest method appeared to lag.

Under MNAR, where the scenario led to African American and Hispanic patients who died were more likely to have missing racial-ethnic group information, relative bias and coverage probabilities were extremely poor for those groups by all imputation methods. In contrast, performance statistics were substantially better for imputed values for mean A1c and mean MPR. This difference may be due to the complex MNAR mechanism for the race groups, which also involved the outcome. It may also be due in part to the data structure: the two racial groups were small compared to the reference group, and there was thus less information available in the data for making effective imputations. For the mean A1c and mean MPR variables there was far more information available even with 50 percent of observations missing, and the imputation algorithms appeared to more effective.

The simulation demonstrated the challenges for applying MNAR sensitivity analysis, even in the unusual situation where the exact missingness mechanism was known. When the pattern mixture model was used to attempt to minimize the relative bias, the imputed datasets contained unrealistic race-ethnicity distributions. Other sensitivity analyses were less successful. In particular, the selection model weighted-parameter method failed because the distribution for the MNAR parameter estimates fell

outside the distribution for the MAR parameter estimates. Carpenter et al. discuss this limitation of the weighting method [36] .

Many of the conclusions drawn from the simulation were repeated with the TBI example. The multiple imputation methods produced relatively unbiased results for the non-Hispanic black group, for which missingness was generally MAR. On the other hand, in the Hispanic group, missingness was MNAR and biased results were seen as a result. In the TBI sensitivity analyses, though the pattern mixture model in which relative bias was minimized (PMM-1) appeared to be generally successful, the Hispanic group in the imputed results was 3.5 times larger than actual. In the pattern mixture model for which the goal was to match the true race distribution within the imputed data (PMM-2), biased results were still seen.

Further work to better understand the MNAR mechanisms that led to the missing race data could help inform future MNAR sensitivity analyses; however, there may be limits to how much can be achieved given the challenges seen in the simulation, where the MNAR situation was fully described.

6. Summary and Conclusions

6.1 Summary

This work was motivated by two challenges that are commonly experienced by investigators who work with large VHA administrative healthcare datasets. These challenges included the need for better ways to account for the patient's disease burden based on diagnostic codes, and the need for improved ways to handle missing data, particularly when the missingness exists in important covariates.

In the first manuscript, improved models for summarizing a patient's disease burden were developed by applying seven machine learning and statistical methods. Each method provided more accurate predictions than models based on the Elixhauser index, and the pooled model, based on the combined predictions of the other six methods, usually had the best predictive performance.

In the second manuscript, an improved comorbidity summary score was developed based on the variable importance measures from the top performing models in the first manuscript. Three large VHA cohorts were used to both train these models and to validate the score. When compared against models based on the Elixhauser index, the score demonstrated more accurate predictive performance.

In the third manuscript, four multiple imputation methods were compared using simulations and applications to real data under several types of missingness. The effectiveness of MNAR sensitivity analyses based on pattern mixture models and selection models was carefully examined, with implications for other VHA investigators who work with similar datasets.

6.2 Discussion and Conclusions

The following paragraphs provide a summary of the insights and conclusions drawn from this research:

6.2.1 Comparing machine learning to traditional statistical methods.

One goal in the first and third aims was to examine whether machine learning methods offered any advantages over traditional statistical approaches, particularly in their ability to automatically account for complex interactions and non-linear effects.

In the first aim, where the goal was to develop better ways to account for disease burden, predictive performance was compared between three machine learning and three statistical methods. The top performers (excluding the pooled model) included two machine learning methods (random forest and Bayesian additive regression trees) and one statistical method (elastic-net penalized logistic regression).

In the third aim, where several multiple imputation methods were compared, machine learning algorithms were incorporated in three models, while the fourth relied on statistical methods for imputation. Here, the top performer in many situations was the

model with random forest incorporated within the multiple imputation with chained equations (MICE) framework, but MICE models using logistic regression and predictive mean matching often achieved similar results. Random forest was far less successful as a stand-alone multiple imputation method, and neural net imputation within the MICE framework also performed poorly.

Overall, neither machine learning nor traditional statistical methods offered a clear advantage over the other group in these applications, and the investigator should carefully consider a wide variety of methods that are not limited to any particular type.

6.2.2 Problems with modeling the correlation structure inherent in the ICD hierarchy.

One goal in the first aim was to take advantage of the hierarchy established by the ICD system, such that if data were sparse for a particular ICD code, information from similar codes within the same hierarchy could be used to approximate the effects for the sparse predictors. This approach was incorporated in the Probability Based Features models (dropped prior to completion of manuscript 1) and Modeled Averaged Regression Coefficients models. Both were among the weakest performing methods, and the original assumption is likely false that the correlation structure imposed by the ICD hierarchy can be used to make valid assumptions about sparsely populated ICD conditions.

6.2.3 Performance advantages of ensemble models.

In the first aim, the model based on the pooled predictions of the other models had the strongest predictive performance of any model. Dietterich [19] described why an

ensemble of accurate and diverse classifiers is likely to perform better than the individual models. A somewhat different ensemble approach was used in the second aim, when the variable importance measures for the top three models applied to four populations were combined such those predictors with importance measures falling in the top 50% in all 12 results were selected for use in the summary score.

6.2.4 New population insights based on variable importance models.

The variable importance results developed in the first and second aims demonstrated that a number of conditions not included in the Elixhauser index were highly predictive for five year mortality. As discussed in the second manuscript, Elixhauser developed the index in order to predict short-term events including in-hospital mortality, hospital charges, or length of stay, and she thus excluded a wide range of conditions from her index [2] since they were not associated with short-term events. It thus was not surprising to find a number of conditions associated with the patient's functional status were strongly associated with longer-term outcomes such as five year mortality. These included Alzheimer's disease, senile dementia, hearing loss, persistent mental disorders, memory loss, falls, lack of household assistance, and urinary incontinence.

6.2.5 Summary score weights based on statistical rather than clinical importance.

The variable importance results in the second aim highlighted that ICD codes for potentially harmful conditions are not always associated with mortality. This is often

related to the codes' primary use as billing mechanisms, such that codes for less serious conditions are often not recorded when the patient is critically ill since other conditions are more likely to be the main drivers of the patient's medical costs. As a result, it would be possible to falsely conclude from ICD data that high blood pressure is protective against mortality [2]. Rather than exclude these associations, which are accurate from a statistical view, codes that predicted survival were instead included in summary score with a negative weighting, even when this appeared to contradict clinical evidence. As a result, many patients had an overall negative score, meaning more of their highly predictive ICD codes were associated with survival rather than mortality. Failure to take advantage of this artifact related to the ICD billing system would lead to substantially worse predictive performance. The disadvantage of such an approach is that the summary score does not provide a clinical picture of the patient's comorbidities; instead, it provides a score used to predict mortality based on a comparison with millions of other Veterans for whether the patient has specific conditions that were most predictive in this population.

6.2.6 Unbiased imputation of continuous variables under MNAR

In the third aim, MNAR scenarios were simulated for two continuous variables and for two of the four levels in a nominal categorical variable (race/ethnicity). None of the multiple imputation methods could provide unbiased results for the categorical variable, but several methods (random forest within MICE and logistic regression/predictive mean matching within MICE) provided unbiased results with slightly narrower confidence intervals than complete case analysis, even when 50% of

values were missing. This difference may be due to the complex MNAR mechanism for the race groups, which also involved the outcome. It may also be due in part to the data structure: the two racial ethnic groups were small compared to the reference group, and there was thus less information available in the data for making effective imputations. For the mean A1c and mean MPR variables there was far more information available even with 50 percent of observations missing, and the algorithms were more effective. In summary, unbiased imputation under MNAR may be possible in some situations, but the investigator must be careful to conduct sensitivity analysis to try to verify the results are reasonable.

6.2.7 Challenges for conducting sensitivity analyses under MNAR

Two types of sensitivity analysis were performed on imputed values in the third aim; these analyses were based on pattern mixture models and selection models (see section 2.3.9). In both the simulations and the application to real data, the MNAR mechanism was well understood. For the real data, this unusual situation existed because different sources of race/ethnicity information became available after the initial cohort had been formed, such that a more accurate variable with substantially lower missingness could be determined for comparison against the original. Thus, sensitivity analysis could be applied in situations where “true” parameter estimates existed. While the pattern mixture model approach could be used to produce reasonable inference, the imputed datasets under those conditions contained unrealistic race-ethnicity distributions. The selection model weighted-parameter method failed because the distribution for the MNAR parameter estimates fell outside the distribution for the MAR

parameter estimates. While this work did not provide a solution for the MNAR race/ethnicity challenge likely faced by many investigators, it did provide important insights into the specific challenges researchers face when the reason for missingness is related to the missing data itself, or to other unknown variables.

6.3 Limitations

Three Veteran populations were studied, with an average age of 73.7, and with an average of 4.2% women. Further work is needed to determine if the results from the first two aims could be generalized to a wider population. The results of the third aim involving missing data are less likely to be affected by the distinct populations.

While each aim considered a wide variety of available methods, these were limited to those which could be completed in a reasonable time on a typical desktop computer (64 bit machine with 16GB RAM, 2.56GHz processor) or on a shared server (64 bit server with 16GB RAM and a 2.36 GHz processor). While this limitation helped to ensure these methods can be directly applied by most investigators, additional methods could be attempted in a parallel environment.

6.4 Future Work

Numerous areas for additional work were noted. Since the original dissertation aims were developed to support specific research problems encountered in work with VHA administrative healthcare data, these future work goals are also well suited for application to VHA research.

1) In the second aim, the summary score algorithm involved a simple weighting scheme, and this was validated only on Veteran populations. Additional work is planned in the following areas:

a. The score will be validated on other groups, including non-Veteran populations to help determine generalizability beyond the older, male population in which the score was originally developed. Other weighting schemes could boost predictive performance, and validation in other outcomes such as one-year mortality could widen its applicability.

b. The score's definition will be expanded to include other types of administrative data, including vital signs, health services utilization, medications, and laboratory tests using a similar approach to that developed for the first two aims; a wide variety of methods will be examined for each of the data types, and competing variable importance measures will be used to identify a group of variables with the strongest predictive performance. This expanded score will be compared to existing measures, such as the Care Assessment Needs Score [80].

c. Since the score is similar in some ways to a propensity score [81], applications for its use in adjusting for confounding and selection bias will be examined. For example, in studies that examine disparities in health outcomes, the comorbidity

score could be used to match patients by various exposures to attempt to control for selection bias or confounding.

2) Expanding further on the third aim, the following additional areas concerning multiple imputation merit further work:

a) Other multiple imputation methods could be applied to the missing race-ethnicity problem for comparison against the existing results. For example, Gebregziabher and DeSantis [82] and Vermunt et al. [83] apply latent class models within the multiple imputation framework. Such models are typically limited to only categorical data, and it would be difficult to apply them directly to most administrative healthcare datasets. However, LCMI may still be a useful tool for investigating the missing race-ethnicity problem.

b) Bayesian Additive Regression Trees (BART) could be tested as the imputation method within the MICE framework within a parallel computing environment. While this method was successfully implemented in the third aim, it was too slow in a single-processor environment to be feasible.

3) The MNAR sensitivity analyses conducted as part of the third aim were limited to missing race-ethnicity, and thus involved only a multi-level, nominal categorical variable. This work will be expanded to include all variable types and a wider range of methods in order to better understand the limitations of current work in this area, and to look for areas where methodological development is warranted. The following paragraphs provide an expanded summary of possible approaches when MNAR conditions are suspected, beyond those already considered in the third aim:

a) The first approach involves developing a joint model that attempts to incorporate the MNAR mechanism. Such a joint distribution may be very complex and inference may require MCMC methods. As Molenberghs and Lesaffre discuss [33], such models are based on untestable assumptions, and may be very sensitive to minor changes in such assumptions. For these reasons sensitivity analyses based on models developed from an MAR basis are more commonly seen.

b) Another option is to consider the addition of auxiliary variables, which may help to explain why missingness occurred, but are otherwise not useful for explaining the outcome. Raykov et al. describe that such variables could be included in the maximum likelihood or multiple imputation models, and could perhaps help the models meet the underlying MAR assumption [84]. However this approach, like the basic MNAR models in paragraph (1), may still rely on untestable assumptions, and it may not be possible to identify the correct auxiliary variables to produce the needed improvement.

c) The most common approach is to conduct sensitivity analyses on MAR models to test whether the inference from models is sensitive to the imposed changes. The analyst must select what types of analyses to conduct based on a “best guess” for what caused the missingness. As discussed in the third aim, most sensitivity analyses are broadly grouped into selection model or pattern mixture model approaches [32, 85]. The data type will further dictate how the analyses are conducted.

i) Sensitivity analyses based on pattern mixture models typically involve one or more shift parameters. For continuous variables, scale factors and shift factors are multiplied by or added to the imputation results. For categorical data, shift

parameters are applied within the appropriate generalized linear model; for example, in logistic regression, the shift represents the change in the log-odds that a specific level of the variable is observed. Different shifts could be imposed for each level of the variable [86].

ii) Selection model sensitivity analyses can take numerous forms; one method was examined in the third aim [36]. Another approach involves developing a measure of local influence. Here the goal is to produce an index that quantifies how much an MAR model deviates from its MLE when it is perturbed towards an MNAR condition. Verbeke et al. [30] derived this approach for normally-distributed longitudinal data, and Troxel et al. demonstrated a similar method for generalized linear models [87]. While Verbeke's approach examines the model's behavior at the individual level, Troxel is concerned with behavior at the group level. Troxel's Index of Sensitivity to non-ignorability (ISNI) is easily implemented since it relies only on determining the MLE from complete case data, and a separate model for predicting missingness.

d) Summary and description of future work: while much work is available in the literature concerning the development of MNAR models (paragraphs a and b above), there is a strong consensus that this effort is less likely to be successful because the basic assumptions for such models are untestable. Sensitivity analyses continue to offer the most promise, and my work will focus on two areas:

i) Pattern mixture model adjustments for other variable types besides nominal categorical data, particularly continuous variables.

ii) Local influence analyses, particularly as described by Troxel [86].

APPENDIX A: EXAMPLES OF R AND SAS PROGRAMS DEVELOPED TO SUPPORT RESEARCH AIMS

A.1 Introduction

This Appendix provides a description of the software program files that were developed to support this research. These files and supporting sample datasets are available in GitHub (user: rccward, repository: comorbidity-models). All R functions described in this appendix were developed using R version 3.2.3 [88]. SAS macros were developed using SAS software (SAS Institute Inc., Cary, NC), version 9.4.

A.2 Comorbidity models development (first manuscript)

A.2.1 Description (models_func)

This R function (models_func) applies the top-performing methods applied in the first aim to summarize disease burden from ICD-9-CM data by training and validating models and comparing each method's predictive performance with models based on the Elixhauser index using AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and net reclassification improvement statistics for events and non-events. Methods include elastic-net regularized generalized linear

model, random forest, Bayesian additive regression trees, and an ensemble model based on the pooled predictions of the other models.

A.2.2 Usage (models_func)

```
models_func(patient_dat, binary_dat, elixmat, iterations, size, covariate_flag)
```

A.2.3 Arguments (models_func)

patient_dat	Dataframe containing the outcome and covariates such as race, age, gender, marital status, and clinical variables. Categorical variables should be stored as factors
binary_dat	Dataframe containing binary ICD-9 data, one row per patient and one column per unique ICD-9 code. Variables should be stored as factors. The column names should list each 5 digit code.
elixmat	Dataframe containing the Elixhauser-Quan comorbidities, one row per patient, and one binary column for each of the 31 comorbidities, stored as factors. Alternatively, the user could establish a different comparison comorbidity measure in place of the Elixhauser index, where each column serves as an independent predictor in the comparison model.
iterations	The number of bootstrapped training and validation samples to be generated in order to determine the distribution of the comparison statistics.
size	The number of observations within each bootstrapped sample.

`covar_flag` A flag indicating whether patient covariates from `patient_dat` should be included in the models, or whether inference should be based solely on ICD-9 information.

A.2.4 Output objects (`models_func`)

A single list is returned containing the following objects:

`meanvarimpRF` A vector of the random forest ICD-9 variable importance measures, with one measure for each column of `binary_mat`.

`meanvarimpBART` A vector of the Bayesian additive regression trees ICD-9 variable importance measures, with one measure for each column of `binary_mat`.

`meanvarimpREG` A vector of the elastic-net regression ICD-9 variable importance measures, with one measure for each column of `binary_mat`.

`output` A matrix containing the performance statistics and their 95% confidence limits for models based on the Elixhauser index, random forest, BART, elastic-net, and the pooled model. Statistics include AUC, sensitivity, specificity, Brier score, net reclassification error (NRI) for events and non-events.

`train_id, test_id` Indices of training and validation observations used in partitioning the dataset such that the same partition is used during summary score development (`score_fn`, below).

A.2.5 Examples (models_func)

Sample datasets `patient_dat`, `binary_dat`, and `elixmat` are provided with the program code in GitHub. These are simulated observations for a diabetes patient population similar to the Veteran population studied in aims 1 – 3. The following console summary is provided in addition to the information stored in the returned object:

```
pred_out<-models_func(dat,binary, elixmat,iterations, size, covar_flag)
```

```
Model Performance Comparison (covariates not included in model)
```

```
1000 iterations
```

```
size = 2000 patients
```

	Elix	RF	BART	REG	Pool
AUC-UCL	0.709	0.834	0.841	0.840	0.854
AUC	0.684	0.820	0.823	0.823	0.837
AUC-LCL	0.663	0.799	0.806	0.803	0.819
sens-UCL	0.537	0.702	0.748	0.709	0.733
sens	0.485	0.672	0.719	0.663	0.701
sens-LCL	0.424	0.627	0.680	0.609	0.660
spec-UCL	0.813	0.841	0.802	0.842	0.834
spec	0.772	0.818	0.774	0.820	0.810
spec-LCL	0.731	0.788	0.739	0.794	0.785
Brier-UCL	0.228	0.182	0.178	0.180	0.173
Brier	0.222	0.172	0.171	0.171	0.164
Brier-LCL	0.215	0.164	0.161	0.162	0.155
NRInevent-UCL	-----	0.246	0.280	0.225	0.273
NRInevent	-----	0.188	0.234	0.178	0.217
NRInevent-LCL	-----	0.142	0.189	0.119	0.169
NRInonevent-UCL	-----	0.088	0.041	0.095	0.079
NRInonevent	-----	0.045	0.002	0.048	0.038
NRInonevent-LCL	-----	0.018	-0.030	0.016	0.008

A.3 Summary score development (second manuscript)

A.3.1 Description (score_fn)

This function demonstrates the method used in the second aim to develop a summary score based on variable importance measures from the top performing models in aim 1. Those ICD-9 codes with importance measures in the top 50% among all models were included in the summary score, and codes associated with mortality and survival were assigned weights of +1 or -1, respectively. The patient's score is a simple weighted sum of how many of the selected ICD-9 codes were found in the patient's record. Score performance was compared to models based on the Elixhauser-Quan index using AUC, sensitivity, specificity, Brier Index, and net reclassification index statistics. Note that the models function (models_func) described above must be run first since score_fn requires variable importance rankings from models_func in order to develop the summary score.

A.3.2 Usage (score_fn)

```
Score_fn(models_out, patient_dat, binary_dat, elixmat, iterations, size, covar_flag)
```

A.3.3 Arguments (score_fn)

models_out	This is the list object produced by models_func, above.
patient_dat	Dataframe containing the outcome and covariates such as race, age, gender, marital status, and clinical variables. Categorical variables should be stored as factors.

binary_dat	Dataframe containing binary ICD-9 data, one row per patient and one column per unique ICD-9 code. Variables should be stored as factors. The column names should list each 5 digit code.
elixmat	Dataframe containing the Elixhauser-Quan comorbidities, one row per patient, and one binary column for each of the 31 comorbidities, stored as factors. Alternatively, the user could establish a different comparison comorbidity measure in place of the Elixhauser index, where each column serves as an independent predictor in the comparison model.
iterations	The number of bootstrapped training and validation samples to be generated in order to determine the distribution of the comparison statistics.
size	The number of observations within each bootstrapped sample.
covar_flag	A flag indicating whether patient covariates from patient_dat should be included in the models, or whether inference should be based solely on ICD-9 information.

A.3.4 Output objects (score_fn)

Score_fn produces a list with the following objects:

comorbidities	This is a list of the ICD-9 codes from binary_mat which were included in the summary score.
weights	This is the weights (+1 for mortality, -1 for survival) assigned to each code in the score.

output Matrix of performance statistics and 95% confidence limits for the model based on the Elixhauser-Quan index and the summary score.

A.3.5 Examples (score_fn)

Console output:

```
>score_run<-score_fn(pred_out,dat,binary, elixmat, iterations,size, covar_flag)
```

Score performance (covariates not included in model)

1000 iterations

	Elix	Summary Score
AUC-UCL	0.693	0.827
AUC	0.676	0.816
AUC-LCL	0.663	0.806
sens-UCL	0.512	0.719
sens	0.472	0.700
sens-LCL	0.436	0.681
spec-UCL	0.795	0.796
spec	0.773	0.781
spec-LCL	0.744	0.765
Brier-UCL	0.228	0.179
Brier	0.224	0.174
Brier-LCL	0.219	0.169
NRInevent-UCL	-----	0.265
NRInevent	-----	0.228
NRInevent-LCL	-----	0.184
NRInonevent-UCL	-----	0.038
NRInonevent	-----	0.008
NRInonevent-LCL	-----	-0.024

A.4 Missing data analyses simulation (third aim)

A.4.1 Description (`missdat_sim`)

This function compares several machine learning and model-based multiple imputation methods for dealing with missing covariate data under missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) scenarios. Beginning with complete case data, the function simulates the desired missingness scenario, and imputation performance is evaluated using relative bias, root mean squared error, efficiency and coverage probability statistics.

A.4.2 Usage (`missdat_sim`)

```
Missdat_sim(patient_dat, MissType, pctMiss, size, iterations)
```

A.4.3 Arguments (`missdat_sim`)

<code>patient_dat</code>	Dataframe containing the outcome and covariates such as race, age, gender, marital status, and clinical variables. Categorical variables should be stored as factors.
<code>MissType</code>	The missingness pattern to be simulated in <code>patient_dat</code> . The options are “MCAR”, “MAR”, “MNAR”. See details (below) for further information concerning the missingness scenarios.
<code>pctMiss</code>	The fraction of patients in the bootstrapped dataset with any missing value. Options are restricted to .1, .3, and .5.

iterations	The number of bootstrapped training and validation samples to be generated in order to determine the distribution of the comparison statistics.
size	The number of observations within each bootstrapped sample.

A.4.4 Details (misssdat_sim)

In the example dataset, under MCAR, missingness is generated for race, mean_mpr and mean_A1c variables completely at random. Under MAR, the probability of missing race is higher when the patient died, the probability of missing mean_mpr is higher when the patient is not married, and the probability of missing mean_A1c is greater when the patient lives in a rural area. Under MNAR, missing race is more likely when the patient is non-Hispanic black or Hispanic and died; missing mean_mpr is more likely when the patient's medication possession ratio (MPR) is above the median value among all patients; missing mean_A1c is more likely when the patient's mean A1c level is above the median value for all patients.

A.4.5 Output objects (misssdat_sim)

misssdat_sim returns the following objects within a single list:

output	This is a list of dataframes, one per multiple imputation method. Each provides the performance statistics with 95% confidence limits for that method, including relative bias, efficiency, root mean squared error, and coverage probability.
--------	--

prob.miss This provides a list of matrices, one per iteration, that summarizes the missingness probabilities under each scenario. This permits the user to verify that the requested scenario was generated.

odd.miss This provides a list of matrices, one per iteration, that summarizes the missingness odds ratios for each scenario. Odds ratios are generated from separate logistic models for each variable with missing values, where the outcome is a missingness indicator. This permits the user to verify that the requested scenario was generated.

A.4.6 Example (missdat_sim)

R console output example:

```
> MissType="MNAR"
> pctMiss=.3
> Nobs=1000
> iterations=1000
> missrun<-missdat_sim(dat,MissType,pctMiss,Nobs,iterations)

-----
> missrun[[1]][[1]]
      misstype  pctMiss  MIType  intercept  single1  rural1
rel.bias-median      3      2      1 -0.08798639 -0.01950551  0.2395583
rel.bias-UCL          3      2      1  0.14671663  0.91969042  3.3358113
rel.bias-LCL          3      2      1 -0.34776225 -1.28024266 -1.5368577
ratio_var-median     3      2      1  43.06712909  43.30529920  41.1828111
ratio_var-UCL        3      2      1  46.45898708  46.09625178  42.1956729
ratio_var-LCL        3      2      1  3.00000000  2.00000000  1.0000000
rmse-median          3      2      1  1.67613116  0.31493434  0.2080584
rmse-UCL             3      2      1  2.71125326  0.40489587  0.3515986
rmse-LCL             3      2      1  3.00000000  2.00000000  1.0000000
CovProb-pct         3      2      1  1.00000000  0.66666667  0.6666667
      male1      age      race2      race3      race4
rel.bias-median  0.1954343 -0.05470550 -14.729965 -6.3138931 -0.1648961
rel.bias-UCL     0.4009245  0.23325525 -10.182891 -4.6691325 -0.1515784
rel.bias-LCL    -1.4137884 -0.12348141 -16.919479 -7.7017439 -0.2718389
ratio_var-median 39.0422326 41.81670229 199.818772 156.8015798 45.6921451
ratio_var-UCL   44.2419931 44.26315920 434.076280 484.9759815 59.3085076
ratio_var-LCL   -0.3477623 -1.28024266 -1.536858 -1.4137884 -0.1234814
rmse-median     0.6738387  0.01871629  2.659716  2.1834179  0.3735142
rmse-UCL        0.8181008  0.02679147  3.095889  2.7849435  0.4686748
rmse-LCL        3.0000000  2.00000000  1.000000  -0.3477623 -1.2802427
CovProb-pct     1.0000000  1.00000000  0.000000  0.0000000  1.0000000
```

	mean1c	meanmpr
rel.bias-median	-1.1485749	-0.06151144
rel.bias-UCL	-0.6711596	0.07616864
rel.bias-LCL	-1.2526204	-0.17614600
ratio_var-median	53.8702350	52.63004672
ratio_var-UCL	65.4548691	55.82645845
ratio_var-LCL	-16.9194787	-7.70174393
rmse-median	0.1643232	0.44817799
rmse-UCL	0.1722005	0.63815386
rmse-LCL	-1.5368577	-1.41378845
CovProb-pct	0.6666667	1.00000000

A.5 Sensitivity analyses: selection model

A.5.1 Description

This simulation program written in R implements the method described by Carpenter et al. for selection model sensitivity analysis after multiple imputation under MAR [36]. The program in its current form requires the use of the example dataset, `patient_dat`, which is stored with the program on GitHub.

A.5.2 Usage

Given the example dataset, a separate function is used to generate MNAR missingness in the race/ethnicity variable in which race is more likely to be missing in non-Hispanic blacks and Hispanics who died. These data are then imputed under MAR assumptions using MICE with logistic regression imputation [26]. Next, the weighted sensitivity approach is applied through an iterative process to examine candidate values for the delta vector, which adjusts how strongly a given level of the race variable is associated with the logodds that it is missing [36]. The user then selects the best values for delta from these results.

A.5.3 Arguments

The following arguments must be provided:

patient_dat	Dataframe containing the outcome and covariates such as race, age, gender, marital status, and clinical variables. Categorical variables should be stored as factors.
MissType	Must be set to “MNAR” for the purposes of the sensitivity analysis.
pctMiss	The fraction of patients in the bootstrapped dataset with any missing value. Options are restricted to 0.1, 0.3, and 0.5. Default value is 0.3.
delta	Elements of this vector adjust how strongly a given level of the race variable is associated with the logodds that it is missing. The user could optionally adjust the coded values that are iteratively tested.

A.5.4 Output objects

Output	An output matrix provides a summary of results for each combination of delta values. These include the estimated race coefficients under MNAR, with relative bias, efficiency, root mean squared error and coverage probability provided for each. Finally, the mortality odds ratios for the MNAR coefficient estimates are provided.
--------	--

Weights This is a list, with a separate matrix for each iteration, providing the calculated weights. See Carpenter et al. [36] for further information on the weights.

A.6 Sensitivity analyses: pattern mixture model

A.6.1 Description

This SAS program demonstrates the method for adjusting each level of the imputed race/ethnicity variables using pattern mixture models [32]. For now, it requires the use of the provided example dataset.

A.6.2 Usage

The demonstration includes three macros which work together:

`%macro missgen` This macro generates MNAR missingness for the race variable in the provided SAS dataset (simdat). See sections A.4 and A.5 for more information on the MNAR association. The macro requires a separately provided SAS program file (missgen.sas) be available. The user should modify the first line of the macro to indicate where this file is stored. The SAS dataset simdat must be loaded in the user's SAS work directory.

`%macro tune` This macro uses SAS PROC MI to perform multiple imputations with pattern mixture model adjustments. Four inputs are required (variables zero, one, two, three); these

are the adjustments to each of the race/ethnicity variables. PROC GENMOD and PROC MIANALYZE are then used to analyze and combine the multiple imputation results.

`%macro shell`

This macro helps the user to iteratively determine the best combination of the four race/ethnicity adjustment parameters. For the range of selected values and for the total number of desired iterations, each combination of adjustments is tested using the `%missgen` and `%tune` macros.

A.6.4 Output objects

The results of `%shell` are stored in two files found in the work directory:

`results_freq`

This table provides the race distribution for each imputed dataset after MNAR adjustment.

`results_OR`

This table provides the results for each combination of the adjustment parameters. Provided results include parameter estimates and their standard errors, the mortality odds ratio and 95% confidence limits, relative bias, efficiency, root mean squared error, and coverage probability.

References Cited

- [1] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.," *J. Chronic Dis.*, vol. 40, no. 5, pp. 373–83, 1987.
- [2] A. Elixhauser, C. Steiner, D. Harris, and R. Coffey, "Comorbidity measures for use with administrative data," *Med. Care*, vol. 36, pp. 8–27, 1998.
- [3] M. Mor, "Assessing Race and Ethnicity," in *VIReC Database & Methods Cyberseminar Series*, 2016, pp. 1–54.
- [4] J. A. Long, M. I. Bamba, B. Ling, and J. A. Shea, "Missing Race/Ethnicity Data in Veterans Health Administration Based Disparities Research: A Systematic Review," *J. Health Care Poor Underserved*, vol. 17, no. 1, pp. 128–140, 2006.
- [5] N. R. Kressin, B. H. Chang, A. Hendricks, and L. E. Kazis, "Agreement between Administrative Data and Patients' Self-Reports of Race/Ethnicity," *Am. J. Public Health*, vol. 93, no. 10, pp. 1734–1739, 2003.
- [6] M. Gebregziabher and Y. Z. Hao, "Lessons learned in dealing with missing race data: an empirical investigation," *J. Biom. Biostat.*, vol. 3, no. 3, p. 3, 2012.
- [7] R. Deyo, D. Cherkin, and M. Ciol, "Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases," *J. Clin. Epidemiol.*, vol. 45, pp. 613–619, 1992.
- [8] P. S. Romano, L. L. Roost, and J. G. Jollis, "Presentation adapting a clinical comorbidity index for use with ICD-9-CM administrative data: Differing perspectives," *J. Clin. Epidemiol.*, vol. 46, no. 10, pp. 1075–1079, 1993.
- [9] H. Quan *et al.*, "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.," *Med. Care*, vol. 43, no. 11, pp. 1130–1139, Nov. 2005.
- [10] Y. T. Chu, Y. Y. Ng, and S. chi Wu, "Comparison of different comorbidity measures for use with administrative data in predicting short- and long-term mortality.," *BMC Health Serv. Res.*, vol. 10, no. 1, p. 140, 2010.
- [11] C. van Walraven, P. C. Austin, A. Jennings, H. Quan, and A. J. Forster, "A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data.," *Med. Care*, vol. 47, no. 6, pp. 626–633, 2009.
- [12] R. E. Kheirbek, F. Alemi, and R. Fletcher, "Heart failure prognosis: comorbidities matter," *J Palliat Med*, vol. 18, no. 5, pp. 447–452, 2015.
- [13] C. Levy *et al.*, "Predictors of 6-Month Mortality among Nursing Home Residents: Diagnoses Maybe More Predictive Than Functional Disability.," *J. Palliat. Med.*, vol. 17, no. X, pp. 1–7, 2014.
- [14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd Ed. Springer, 2008.
- [15] B. J. Wolf, E. G. Hill, and E. H. Slate, "Logic forest: An ensemble classifier for discovering logical combinations of binary markers," *Bioinformatics*, vol. 26, no. 17, pp. 2183–2189, 2010.
- [16] D. J. Stekhoven and P. Buhlmann, "MissForest--non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

- [17] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, "Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study," *Am. J. Epidemiol.*, vol. 179, no. 6, pp. 764–774, 2014.
- [18] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," *J. Stat. Comput. Simul.*, vol. 84, no. 6, pp. 1313–1328, 2014.
- [19] T. G. Dietterich, "Ensemble Methods in Machine Learning," 2000.
- [20] M. J. G. Leening, M. M. Vedder, J. C. M. Witteman, M. J. Pencina, and E. W. Steyerberg, "Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide.," *Ann. Intern. Med.*, vol. 160, no. 2, pp. 122–31, Jan. 2014.
- [21] M. S. Pepe, J. Fan, Z. Feng, T. Gerds, and J. Hilden, "The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets," *Stat. Biosci.*, vol. 7, no. 2, pp. 282–295, 2015.
- [22] G. W. Brier, "Verification of forecasts expersses in terms of probaility.," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
- [23] K. T. Stroupe, E. Tarlov, Q. Zhang, T. Haywood, A. Owens, and D. M. Hynes, "Use of Medicare and DOD data for improving VA race data quality.," *J. Rehabil. Res. Dev.*, vol. 47, no. 8, pp. 781–795, 2010.
- [24] M. W. Sohn *et al.*, "Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs.," *Popul. Health Metr.*, vol. 4, p. 7, 2006.
- [25] D. B. Rubin, "An Overview of Multiple Imputation," *Proc. Surv. Res. methods Sect. Am. Stat. Assoc.*, pp. 79–84, 1988.
- [26] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.
- [27] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Stat. Med.*, vol. 30, no. 4, pp. 377–399, 2011.
- [28] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [29] J. M. Jerez *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–115, 2010.
- [30] G. Verbeke, G. Molenberghs, H. Thijs, E. Lesaffre, and M. G. Kenward, "Sensitivity Analysis for Nonrandom Dropout : A Local Influence Approach," *Biometrics*, vol. 57, no. 1, pp. 7–14, 2001.
- [31] G. Molenberghs, C. Beunckens, C. Sotito, and M. Kenward, "Every missingness not at random model has a missingness at random counterpart with equal fit," *J.R.Statist. Soc. B*, vol. 70, no. 2, pp. 371–388, 2008.
- [32] J. Carpenter and M. Kenward, *Multiple Imputation and its Application*. West Sussex, UK: John Wiley & Sons, 2013.
- [33] G. Molenberghs and E. Lesaffre, "Missing data," *BMJ*, vol. 334, p. 424, Feb. 2007.
- [34] H. Thijs, G. Molenberghs, B. Michiels, G. Verbeke, and D. Curran, "Strategies to fit pattern-mixture models," *Biostatistics*, vol. 3, no. 2, pp. 245–265, 2002.
- [35] P. Diggle and M. G. Kenward, "Informative Drop-Out in Longitudinal Data Analysis," *J. R. Stat. Soc. Ser. C (Applied Stat.)*, vol. 43, no. 1, pp. 49–93, 1994.

- [36] J. R. Carpenter, M. G. Kenward, and I. R. White, "Sensitivity analysis after multiple imputation under missing at random: a weighting approach," *Stat. Methods Med. Res.*, vol. 16, no. 3, pp. 259–275, 2007.
- [37] A. Marshall, D. G. Altman, and R. L. Holder, "Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study.," *BMC Med. Res. Methodol.*, vol. 10, no. 1, p. 112, 2010.
- [38] K. I. Vaden, M. Gebregziabher, S. E. Kuchinsky, and M. a. Eckert, "Multiple imputation of missing fMRI data in whole brain analysis," *Neuroimage*, vol. 60, no. 3, pp. 1843–1855, 2012.
- [39] B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj, "Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait--a cohort study.," *BMJ Open*, vol. 3, no. 5, pp. 1–10, 2013.
- [40] P. C. Austin, "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality," *Stat. Med.*, vol. 26, no. 15, pp. 2937–2957, Jul. 2007.
- [41] S. Kennedy, Wiitala, Hayward, "Improved cardiovascular risk prediction using nonparametric regression and electronic health record data," *Med. Care*, vol. 51, no. 3, pp. 251–258, 2013.
- [42] A. Neumann, J. Holstein, J. R. Le Gall, and E. Lepage, "Measuring performance in health care: case-mix adjustment by boosted decision trees," *Artif. Intell. Med.*, vol. 32, no. 2, pp. 97–113, 2004.
- [43] X. Song, A. Mitnitski, J. Cox, and K. Rockwood, "Comparison of machine learning techniques with classical statistical methods in predicting health outcomes," vol. 107, pp. 736–740, 2004.
- [44] J. Wu, J. Roy, and W. F. Stewart, "Prediction Modeling Using EHR Data," *Med. Care*, vol. 48, no. 6, pp. S106–S113, 2010.
- [45] R. A. Deyo, D. C. Cherkin, and M. A. Ciol, "Adapting a Clinical Comorbidity Index for Use with ICD-9-CM Administrative Databases," *J Clin Epidemiol*, vol. 45, no. 6, pp. 613–619, 1992.
- [46] J. Siddique, G. W. Ruhnke, A. Flores, and M. T. Prochaska, "Applying Classification Trees to Hospital Administrative Data to Identify Patients with Lower Gastrointestinal Bleeding," 2015.
- [47] C. P. Lynch, M. Gebregziabher, Y. Zhao, K. J. Hunt, and L. E. Egede, "Impact of medical and psychiatric multi-morbidity on mortality in diabetes: emerging evidence.," *BMC Endocr. Disord.*, vol. 14, no. 1, p. 68, Jan. 2014.
- [48] C. E. Dismuke, M. Gebregziabher, D. Yeager, and L. E. Egede, "Racial / Ethnic Differences in Combat- and Non – Combat-Associated Traumatic Brain Injury Severity in the Veterans Health Administration : 2004 – 2010," *Am J. Public Heal.*, vol. 105, no. 8, pp. 1696–1698, 2015.
- [49] M. Gebregziabher, L. Egede, G. E. Gilbert, K. Hunt, P. J. Nietert, and P. Mauldin, "Fitting parametric random effects models in very large data sets with application to VHA national data.," *BMC Med. Res. Methodol.*, vol. 12, p. 163, Jan. 2012.
- [50] E. Hoerl and R. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Dec. 1970.

- [51] R. Tibshirani, "Regression Shrinkage and Selection via Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Nov. 1996.
- [52] H. Zou and T. Hastie, "Regularization and Variable Selection via the Elastic Net," *JRSSB*, vol. 67, no. 2, pp. 301–320, 2005.
- [53] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.
- [54] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 2008–2010, 2010.
- [55] L. G. Glance, T. M. Osler, D. B. Mukamel, W. Meredith, J. Wagner, and A. W. Dick, "TMPM–ICD9," *Ann. Surg.*, vol. 249, no. 6, pp. 1032–1039, Jun. 2009.
- [56] D. J. Hand and K. Yu, "Idiot's Bayes---Not So Stupid After All?," *Int. Stat. Rev.*, vol. 69, no. 3, pp. 385–398, 2001.
- [57] G. B. Hahsler M, Buchta C., "arules: Mining Association Rules and Frequent Itemsets. R package version 1.4-1," 2016. [Online]. Available: <https://cran.r-project.org/package=arules>.
- [58] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [59] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Ann. Appl. Stat.*, vol. 4, no. 1, pp. 266–298, 2010.
- [60] H. Chipman and R. McCulloch, "BayesTree: Bayesian Additive Regression Trees R Package version 0.3-1.4," 2016.
- [61] E. Chrischilles *et al.*, "Beyond Comorbidity," *Med. Care*, vol. 52, no. 3, pp. S75–S84, Mar. 2014.
- [62] L. C. Yourman, S. J. Lee, M. A. Schonberg, E. W. Widera, and A. K. Smith, "Prognostic indices for older adults: a systematic review.," *JAMA*, vol. 307, no. 2, pp. 182–92, Jan. 2012.
- [63] D. Kallogjeri, S. M. Gaynor, M. L. Piccirillo, R. a Jean, E. L. Spitznagel, and J. F. Piccirillo, "Comparison of comorbidity collection methods.," *J. Am. Coll. Surg.*, vol. 219, no. 2, pp. 245–55, 2014.
- [64] H. Quan, G. a Parsons, and W. a Ghali, "Validity of information on comorbidity derived rom ICD-9-CCM administrative data.," *Med. Care*, vol. 40, no. 8, pp. 675–85, Aug. 2002.
- [65] L. M. Baldwin, C. N. Klabunde, P. Green, W. Barlow, and G. Wright, "In search of the perfect comorbidity measure for use with administrative claims data: does it exist?," *Med. Care*, vol. 44, no. 8, pp. 745–53, 2006.
- [66] H. Y. Lin *et al.*, "Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer.," *J. Hum. Genet.*, vol. 53, no. 9, pp. 802–11, 2008.
- [67] R. K. Scher *et al.*, "Onychomycosis Diagnosis and Management: Perspectives from a Joint Dermatology-Podiatry Roundtable," *J Drugs Dermatol*, vol. 14, no. 9, pp. 1016–1021, 2015.
- [68] D. S. Loo, "Onychomycosis in the elderly: Drug treatment options," *Drugs and Aging*, vol. 24, no. 4, pp. 293–302, 2007.

- [69] U.S. Center for Disease Control and Prevention, "ICD-9-CM Code Conversion Table." [Online]. Available: https://www.cdc.gov/nchs/icd/icd9cm_addenda_guidelines.htm#conversion_table. [Accessed: 01-Jan-2016].
- [70] S. Schneeweiss and M. Maclure, "Use of comorbidity scores for control of confounding in studies using administrative databases.," *Int. J. Epidemiol.*, vol. 29, pp. 891–898, 2000.
- [71] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [72] S. van Buuren and G. K. "mice: Multivariate Imputation by Chained Equations. R package version 2.9," 2011. [Online]. Available: <http://cran.r-project.org/package=mice>.
- [73] R. Little, "Missing Data Adjustments in Large Surveys," *J. Bus. Econ.*, vol. 6, pp. 287–301, 1988.
- [74] B. Kindo, "Package 'mpbart' R package v. 0.2," 2016.
- [75] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth. New York: Springer, 2002.
- [76] M. et al. Kuhn, "R package CARET," 2016.
- [77] G. Molenberghs and M. Kenward, *Missing Data in Clinical Trials*, First. West Sussex, UK, 2007.
- [78] V. Héraud-Bousquet, C. Larsen, J. Carpenter, J. C. Desenclos, and Y. Le Strat, "Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data," *BMC Med. Res. Methodol.*, vol. 12, p. 73, 2012.
- [79] A. Hapfelmeier, T. Hothorn, C. Riediger, and K. Ulm, "Estimation of a Predictor's Importance by Random Forests When There Is Missing Data: RISK Prediction in Liver Surgery using Laboratory Data," *Int. J. Biostat.*, vol. 10, no. 2, pp. 165–183, Jan. 2014.
- [80] L. Wang *et al.*, "Predicting Risk of Hospitalization or Death Among Patients Receiving Primary Care in the Veterans Health Administration table 2," vol. 51, no. 4, pp. 368–373, 2013.
- [81] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
- [82] M. Gebregziabher and S. M. DeSantis, "Latent class based multiple imputation approach for missing categorical data," *J. Stat. Plan. Inference*, vol. 140, pp. 3252–3262, 2010.
- [83] J. K. Vermunt, J. R. van Ginkel, L. A. van der Ark, and K. Sijtsma, "Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis," *Sociol. Methodol.*, vol. 38, no. 1, pp. 369–397, Aug. 2008.
- [84] T. Raykov and G. a. Marcoulides, "Identifying Useful Auxiliary Variables for Incomplete Data Analyses: A Note on a Group Difference Examination Approach," *Educ. Psychol. Meas.*, vol. 74, no. 3, pp. 537–550, 2013.
- [85] D. Curran, G. Molenberghs, H. Thijs, and G. Verbeke, "Sensitivity analysis for pattern mixture models.," *J. Biopharm. Stat.*, vol. 14, no. 1, pp. 125–143, 2004.

- [86] Y. Yuan, "Sensitivity Analysis in Multiple Imputation for Missing Data," *SAS Inst. Inc*, pp. 1–12, 2014.
- [87] A. B. Troxel, G. Ma, and D. F. Heitjan, "An index of local sensitivity to nonignorability," *Stat Sin*, vol. 14, no. 4, pp. 1221–1237, 2004.
- [88] "R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>." 2015.