Comparing Internal Validation Methods for a Random Forest Prediction Model of Suicide Death

Qinqing Liao

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Committee:

R. Yates Coley, Chair

Noah Simon

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Comparing Internal Validation Methods for a Random Forest Prediction Model of Suicide Death

Qinqing Liao

Chair of the Supervisory Committee:
Assistant Investigator, R. Yates Coley

Kaiser Permanente Washington Research Institute

Predictive models estimated with clinical data are increasingly popular in the medical data field. After developing a prediction model, its necessary to evaluate its performance in practice, or validate the model. Model validation methods include both internal and external validation; this thesis will focus on the comparison of internal validation methods using a split sample and an entire sample approach. The split sample approach uses a typical randomly selected validation set. For the entire sample approach, we explored three different methods – approximate optimism correction, exact optimism correction and 5-fold cross validation (CV). The dataset included 13,980,570 records on mental health outpatient visits between 2011 - 2017, including information on prior diagnoses, medications, and encounters prior to the visit and follow-up information on suicide death. Data were separated into a development dataset, which

included visits from 2011 - 2014 and was used for model estimation and internal validation, and a prospective validation set, which included visits from 2015 - 2017 and was used to mimic the future data if the model were implemented in clinical practice. We estimated a random forest model to predict suicide death in the 90 days following a visit. We found that the split sample estimation method and 5-fold CV using the entire sample provided more accurate estimation of model performance compared to the exact and optimism correction methods using the entire sample, which both underestimated model optimism and, thus, overestimated model performance in the prospective dataset. Our results stand in contrast to prior research which demonstrated the accuracy of optimism correction methods with logistic regression models estimated using an entire sample approach. While findings may differ for other datasets, model estimation methods, and prediction applications, we recommend caution when using optimism correction methods for internal validation of prediction models estimated in the entire sample when working with very large datasets, rare events, and machine learning prediction models.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

To begin with, I would like to express my sincere gratitude to my advisor R. Yates Coley for the support of my Master study and research, as well as for her patience, motivation, enthusiasm, and immense knowledge. She provided me with help and encouragement all the time during my thesis work as well as my academic study. She also taught me how to be organized, careful and enthusiastic while doing something. I could not have imagined having a better advisor and mentor for my master study.

Besides my advisor, I would like to thank my second reader in the thesis committee: Prof. Noah Simon for his insightful comments, support and hard questions.

My sincere thanks also goes to Gitana Garofalo, for providing me with the warmest support for both of my academic and personal life, as well as organizing various activities in the department to support each other.

Last but not the least, I would like to thank my family: my parents Su Liao and Lihong Qin, for giving birth to me at the first place and supporting me spiritually throughout my life.

Chapter 1. INTRODUCTION

Predictive models, which aim to forecast outcomes given current data, are applied more and more widely in clinical data fields. Various types of models can be used for prediction, ranging from less to more flexible, including generalized linear models, decision trees, random forests and neural networks. After developing a specific prediction model, the next important question is how to estimate its performance in a real-life setting. This process is known as model validation, and it is used to confirm that the output of the model is acceptable with respect to the real-world data situation. There are two different types of model validation: internal and external validation. External validation requires applying a prediction model to an outside data source, and researchers could benefit from it when the dataset used to estimate the original model is small [1]. On the other hand, internal validation estimates prediction model validity using the same data source when an independent dataset is not available or when future performance of the prediction model in the same setting is of interest [7]. This thesis focuses on evaluating different internal validation methods.

Several internal validation methods have been developed and used in previous studies [4][5][8][9]. One defining characteristic of different internal validation methods is whether part of the dataset or the entire dataset is used for model development. For the first "split sample" method, part of the data is randomly selected for model fitting and the remaining part is used for model validation. However, when the number of events is small, estimates of predictive performance using the split-sample method are unstably and may have large bias [5]. In addition, some have argued that split sample approach does not use the data efficiently, since it only uses one part of the data for training and discards the information in the testing sample [5][7][9].

Instead, there are some other validation methods that do not require a separate independent testing set so that all data can be used for model development; these include optimism correction methods and cross validation. For optimism correction, optimism is defined as the difference between the model's true performance and apparent performance – the former refers to performance in the true population whereas the latter indicates the estimated performance in the sample [7]. Optimism correction methods attempt to estimate the optimism from bootstrap samples and then correct overestimation of model performance without an independent testing dataset [3][7]. Meanwhile, cross validation methods use random partitions of the existing data into complementary subsets in order to estimate the model performance in practice. Cross validation processed by using all but one subset for model development, and evaluating performance in the remaining subset. Both of these internal validation methods use the entire dataset for model development and validation minimize the loss of information compared with the split sample method. Using cross validation or optimism correction in the entire sample can also obtain a more stable estimate of model performance.

This thesis examines a setting where the advantages of split sample or entire sample methods for internal validation are not known: a large dataset (millions of records), a rare event (<10 events per 10,000 records), and use of a non-parametric prediction model. Simulation studies have shown little benefit to using entire sample methods in large sample size studies [5][9]. There are also few studies examining the accuracy for these three internal validation methods in a large set with rare events. In addition, existing studies demonstrating the accuracy of internal validation methods used logistic regression to estimate the prediction model [5][9]. While non-parametric methods, e.g., random forest, are becoming more common, there are few studies examining internal validation methods using these models.

The purpose of this thesis is to compare the different internal validation methods for a prediction model to estimate suicide death within 90 days of an outpatient mental health visit. If patients with higher risk of suicide death could be correctly identified using a prediction model, providers may be able to intervene in advance to reduce the risk of suicide. Therefore, this model might be able to provide some clinical guidance for physicians who are dealing with patients with serious mental health problems.

The dataset we used comes from seven health systems, including about 14 million outpatient mental health visits in nine states, and suicide death is a rare event in this data set (2.4 deaths per 10,000 visits) [5]. We used a non-parametric model, random forest, to estimate the prediction model. We examined split sample internal validation as well as multiple entire data estimation approaches: approximate and exact optimism estimation and 5-fold CV estimation.

Chapter 2. METHODS

2.1 DATA DESCRIPTION

Data were collected from seven health systems: Health Partners; Henry Ford Health System; and the Colorado, Hawaii, Northwest, Southern California, and Washington regions of Kaiser Permanente. The analytic dataset includes 13,980,570 outpatient visits to mental health specialty providers for patients age 13 years and older. We split the dataset into two parts, a development dataset (including visits that occurred between January 1, 2011 to September 30, 2014) and a prospective validation dataset (from Oct 1, 2014 to September 30, 2017). The development dataset is used to estimate the prediction model and anticipated model performance with the internal validation methods being compared, and the prospective validation set is used to mimic future new data, as if the prediction model were deployed in a clinical setting, and examine model performance in the new data.

The outcome for our prediction model is suicide death within 90 days following a mental health clinic visit. Since one person may have several mental health visits in the 90 days before a suicide death, the same death may be attributed to multiple visits in the dataset. There are 134 predictors in the analysis data, including demographic characteristics, patient responses to the Patient Health Questionnaire 9-item (PHQ-9), and information on psychiatric medications, diagnoses, prior suicide attempts, and other mental health encounters in the 5 years preceding the visit. For prescription fills, diagnoses, previous suicide attempts, and encounters within the prior 5 years, there are 3 indicator variables representing overlapping time periods: the previous 3 months, previous 1 year, and previous 5 years. The majority of these predictors are indicator variables, as their corresponding values are binary. There are also some categorical variables with more than two levels, e.g. race and ethnicity, as well as some continuous variables, e.g. age and

Charlson score. In addition, there are also some interaction terms, e.g. interaction between age and Charlson score, previous suicide attempts and age. More detailed information for the dataset is listed in the Appendix I.

Predictors in the dataset do not contain any missing values. Indicator variables take the value 0/1, if the value is 0, we cannot distinguish whether the condition was not evaluated (e.g., patient had no mental health care utilization at which depression symptoms were evaluated) or the patient does not satisfy the indicated condition. For categorical variables, e.g. race, the category 'Unknown/Missing' is used for patients with no race recorded in the clinical data. In addition, all predictors are pre-specified before analysis.

2.2 PREDICTION MODEL ESTIMATION AND EVALUATION

We compared two sampling methods for prediction model estimation – split-sample estimation and entire sample estimation. For split-sample estimation, the entire *development dataset* was split into a *training set* containing all visits from a random sample of half the people in model development dataset. All visits from the remaining 50% people comprise the *testing set,* which is used for estimation of model performance. For entire sample estimation, the prediction model was constructed on the entire *development dataset.* In this method, we compare three different estimates of model performance using this *development dataset*: approximate optimism correction, exact optimism correction, and 5-fold cross validation (5-fold CV). Methods for model estimation and performance evaluation for each sampling approach are described below.

Random forests were used to estimate prediction models. Since suicide death is a rare event (24 deaths per 100,000), we used event-stratified bootstrapping to sample visits for trees in the random forest in order to guarantee there are enough events for training each tree.

2.2.1 *Measures of model performance*

Model performance, measured by area under the curve (AUC) and sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) at the 99th, 95th, 90th, and 75th percentiles of the risk score distribution, was evaluated for each estimation methods in the development set and the prospective set.

Some notation for evaluation of model performance are defined below:

Note that for all definitions below, the *risk score threshold* was determined in the training set at each percentile threshold. We then compared this *risk score threshold* to *predicted risk scores* from each prediction model in the testing and prospective validation sets.

TP – true positive, refers to visits which have a suicide death and have a *predicted risk score* higher than the *risk score threshold* at the specified percentile threshold.

FP – false positive, refers to visits which do not have a suicide death but have a *predicted risk score* higher than the *risk score threshold*.

TN – true negative, refers to visits which do not have a suicide death and have a *predicted risk score* lower than the *risk score threshold*.

FN – false negative, refers to visits which have a suicide death but have a *predicted risk score* lower than the *risk score threshold*.

Sensitivity = TP / (TP + FN) x 100%, refers to the percentage of visits with a *predicted risk score* greater than the *risk score threshold* at each percentile threshold among all visits with a suicide death

Specificity = TN / (TN + FP) x 100%, refers to the percentage of visits with a *predicted risk score* less than the *risk score threshold* at each percentile threshold among all visits without a suicide death

PPV = TP / (TP + FP) x 100%, refers to the percentage of visits with suicide death among all visits with a *predicted risk score* higher than the *risk score threshold* at each percentile threshold

NPV = TN / (FN + TN) x 100%, refers to the percentage of visits without suicide death among all visits with a *predicted risk score* lower than the *risk score threshold* at each percentile threshold

2.2.2 *Split-sample prediction model estimation and evaluation*

To begin with, we first examine the model performance using the split sample method for internal validation. We explain the following steps in this section: 1) identify optimal parameters for the random forest model; 2) construct a prediction model with the *training set* using the optimal parameters; and 3) estimate the model performance in the *testing set*.

We began the split sample method by using the *training set* to select two tuning parameters: minimum terminal node size (we considered 10,000, 25,000, 50,000, 100,000, and 150,000) and number of features randomly selected at each node ($2\sqrt{p}$, $\sqrt{p}$, and $0.5\sqrt{p}$, where p is the number of total features in the model), also denoted as *ntry*. The typically recommended value for *ntry* is $\sqrt{p}$, but we also considered $2\sqrt{p}$ and $0.5\sqrt{p}$. It is noted that many predictors in the dataset are correlated, e.g. overlapping time periods indicator variables for each diagnosis. If there are many strong predictors, a smaller value for *ntry* will perform similarly to larger values while reducing computing time. On the other hand, if many predictors are weak, we need to have a larger value for *ntry*.

5-fold CV was used for tuning parameter selection. People in the training set were randomly divided among 5 folds, stratified by people with and without suicide death following

any visit; all visits from a person appeared in the same fold. For a specific parameter combination, a random forest model was constructed on 4 folds and then used to predict the risk of event for the remaining visits in the 5th fold. This process was repeated for each of five folds to get out-of-fold predictions for the entire *training set*. Then, the corresponding out-of-fold AUC was calculated for this parameter combination using these out-of-fold predictions. This 5-fold CV process was repeated for all different parameter combinations. The optimal parameter combination was that with the highest cv-AUC among all tuning parameter combinations considered.

The *prediction model* for the split sample estimation was constructed by estimating a random forest model in the *training set* using the optimal tuning parameters selected by 5-fold CV. This *prediction model* was then used to make predictions for visits in the *validation set*, and the corresponding AUC, sensitivity, specificity, PPV and NPV were calculated. Finally, to obtain 95% confidence intervals (CI) around performance estimates, we re-calculated each measure of test performance in 250 bootstrap samples of validation set visits. The 95% CI was defined as the values at the 2.5th and 97.5th percentile of the bootstrap distributions for each performance measure.

2.2.3 *Entire sample prediction model estimation and evaluation*

Next, we examined model performance using the entire sample method for internal validation. Similarly, we present the following steps in this section: 1) identify optimal parameters for the random forest model; 2) construct a prediction model with the *development dataset* using the optimal parameters; and 3) estimate model performance using the *development dataset.*

For the entire sample approach, we used the entire *development dataset* to select tuning parameters. We used 5-fold CV, similar to the procedure described above for split-sample estimation, but, due to the larger number of visits, we considered larger minimum node sizes (50,000, 100,000, 250,000, and 500,000). We considered the same values ($2\sqrt{p}$, $\sqrt{p}$ and $0.5\sqrt{p}$) for number of features randomly chosen for splitting at each node. Out-of-fold predictions for the *development dataset* with the best out-of-fold AUC were also saved in order to evaluate cross-validated model performance.

The *primary model* for the entire sample estimation method was constructed by estimating a random forest model with the entire *development dataset*, using the optimal parameters chosen from 5-fold CV. The in-sample performance, including AUC, sensitivity, specificity, PPV, and NPV was then calculated for the *primary model* in the *development dataset*.

We used three different internal validation methods to estimate performance of the primary model in a future dataset: approximate optimism correction, exact optimism correction, and cross-validation. To do this, we first obtained event-stratified bootstrap samples of the *development dataset*. We stratified sampling on events to reduce the variability and more efficiently estimate performance. For each bootstrap sample (*sample**), we did the following:

1. Estimated a random forest model (*model**) in *sample** and calculated

    a. the performance (AUC, sensitivity, specificity, PPV, and NPV) of the *model** in the *sample**

    b. the performance of the *model** in the *development dataset*.

    c. the optimism, or difference between a. and b.

2. Calculated the performance of the *primary model* in *sample**, to obtain 95% CI for in-sample performance

3. Calculated the performance of out-of-sample predictions (from 5-fold CV) in *sample**

These estimates from each bootstrap sample are used as input for the validation methods. We stopped bootstrapping the *development dataset* if both of following two criteria were satisfied:

1. We had examined at least 100 bootstrap samples, and
2. The change in the average optimism for all measures of model performance was less than 0.0001

Approximate optimism correction

Estimated prediction model performance with approximate optimism correction was equal to the average performance of prediction models estimated with the bootstrap samples (*model*)* in the *development dataset* (1b above). Quantile-based 95% CIs for each performance measure were defined by the 2.5th and 97.5th percentile of the distribution of the performance of *model** on the *development dataset*.

Exact optimism correction

For the exact optimism correction method, the *optimism* in each bootstrap sample was calculated as the difference between the performance of the bootstrap model in the development dataset and in the bootstrap sample:

Optimism = (performance of *model** in *sample**)

– (performance of *model** in *development dataset*)

Next, the estimated *optimism* was deducted from the performance of the *primary model* in the *development set* to get the *corrected performance*:

corrected performance = (performance of *primary model* in *development dataset*) – *optimism*

The 95% CI for each performance measure was obtained by finding the values at the 2.5th and 97.5th percentile of the bootstrap distributions of corrected performance for each performance measure.

5-fold CV

The out-of-sample predictions for the tuning parameter combination with the best cv-AUC were bootstrapped (as a part of the process described above). We then re-estimated the corresponding AUC, Sensitivity, Specificity, PPV and NPV for each bootstrap sample. The estimation of AUC for this section was the estimated performance of the out-of-sample predictions in the development dataset. The 95% confidence interval for each performance measure is defined by the 2.5 percentile and 97.5 percentile of the estimated performance in the bootstrap sample.

2.2.4 *Model Performance Estimation on Prospective Validation Dataset*

The prediction models estimated from the split sample and entire sample approaches were evaluated in the *prospective validation set* using the same performance measures (AUC, Sensitivity, Specificity, PPV and NPV). In order to get the 95% CIs, each performance measure was calculated in 250 bootstrap samples of the prospective validation set. The 95% CI for each was defined by the 2.5th and 97.5th percentiles of the bootstrap distribution.

We used R (version 3.6.2) programming for all statistical analysis and the ranger package (version 0.12.1) for estimation of random forest models. Other packages can be found in Appendix II. In addition, all code for the analysis is available on github:

https://github.com/qinqing127/Master-Thesis/tree/Coding

There are also some sample codes in Appendix III.

Chapter 3. RESULTS

Table 1 summarizes characteristics of visits in the development and prospective validation datasets. There were 9,610,318 visits in the development dataset and 3,754,137 visits in the prospective validation dataset. In the development dataset, we identified 2,318 suicide deaths, and the observed suicide death rate was 24 cases per 100,000 visits. There were 710 suicide deaths in the prospective validation dataset, and the death rate was 19 cases per 100,000 visits. For both development and prospective validation datasets, visits were most common at the age interval 45-65 years old (36.5% in development dataset and 33.1% in prospective validation dataset) and the majority of visits were by White patients (69.0% in the development dataset and 66.9% in the prospective validation dataset). The percentage of visits with unknown or missing race was around 15% for both the development and prospective validation datasets (15.0% vs 16.7%). For insurance type, most patients had commercial insurance (75.7% and 67.8% in the development and prospective validation datasets, respectively).

Most visits had missing responses for the PHQ-9 9$_{th}$ item, which asks about the frequency of thoughts of suicide or death in the prior 2 weeks in both datasets (91.8% in the development dataset and 74.0% in the prospective validation dataset). We note that clinical implementation of recommending a PHQ-9 at all visits occurred later in the study period, so higher rates of PHQ-9 completion in the prospective validation set, which covers more recent years, was expected. For visits having non-missing responses for the PHQ-9 9$_{th}$ item, most visits had value 0, indicating no days with thoughts of suicide or death (6.1% and 19.8% in the development and prospective validation dataset respectively). Lower rates of suicidal ideation among those completing the PHQ-9 was expected in the prospective validation set since the PHQ-9 was recorded for more patients, rather than being up to provider discretion, which resulted in patients with more severe

symptoms being more likely to have PHQ-9 responses. For diagnoses in the 5 years before a visit, most people in both datasets were diagnosed with depression and anxiety (development dataset:75.5% depression and 70.1% anxiety; prospective validation dataset: 76.0% depression and 79.0% anxiety). Similarly, for prescription fills in past 5 years, antidepressant medication was used most often in both datasets (67.9% in the development dataset and 68.2% in the prospective validation dataset).

A large majority of patients had an outpatient encounter with mental health diagnosis in the 5 years preceding visits (91.5% in the development dataset and 91.4% in the prospective validation dataset). The percentages of previous suicide attempts were similar in both datasets (4.0% in the development dataset and 4.6% in the prospective validation dataset). Finally, for the Charlson comorbidity index, most visits had a score of 0 in both datasets (74.2% in the development dataset and 72.4% in the prospective validation dataset), indicating a low comorbidity burden [2].

Table 1. Characteristics of sampled visit for Development Dataset (January 2009 - August 2014) and Prospective Validation Dataset (January 2015 - September 2017)

	Development dataset (N = 9,610,318 visits)		Prospective validation dataset (N = 3,754,137 visits)	
Characteristic	N	%	N	%
Female	6,081,526	64.1	2,406,695	64.1
Age				
17 or younger	1,048,533	10.9	417,977	11.1
18-29	1,562,616	16.3	670,337	17.9
30-44	2,453,898	25.5	948,716	25.3
45-64	3,510,438	36.5	1,242,067	33.1
65 or older	1,034,813	10.8	475,040	12.7
Race				
White	6,627,544	69.0	2,511,547	66.9
Asian	448,590	4.7	192,284	5.1
Black	836,479	8.7	324,852	8.7
Hawaiian/Pacific Islander	108,410	1.1	32,803	0.9
Native American	92,312	1.0	35,824	1.0
More than one	3,298	0.03	1,788	0.1
Other	48,800	0.5	20,879	0.6
Unknown or Missing	1,444,885	15.0	634,160	16.7
Ethnicity				
Hispanic	2,183,742	22.7	987,061	26.3
Insurance type				
Commercial group	7,275,199	75.7	2,544,926	67.8
Individual	325,139	3.4	141,234	3.8
Medicare	1,409,742	14.7	572,476	15.2
Medicaid	406,606	4.2	413,007	11.0
Other	193,632	2.0	82,494	2.2
PHQ-9 9th item response[1]				
0, Not at all	590,468	6.1	743,288	19.8
1, Several days	121,483	1.3	153,671	4.1
2, More than half the days	43,697	0.5	46,977	1.3
3, Nearly every day	31,408	0.3	32,410	0.9
Missing	8,823,262	91.8	2,777,791	74.0
PHQ-9 recorded in the previous year	1,296,856	13.5	1,623,943	43.3
Diagnoses in past 5 years, including index visit				

Depression	7,257,525	75.5	2,852,694	76.0
Anxiety	6,734,366	70.1	2,967,524	79.0
Bipolar dx	1,317,629	13.7	474,897	12.6
Schizophrenia	400,713	4.2	154,590	4.1
Other psychosis	537,050	5.6	221,351	5.9
Dementia	96,751	1.0	54,782	1.5
Attention deficit disorder	1,151,179	12.0	485,877	12.9
Autism spectrum disorder	131,900	1.4	65,795	1.8
Personality disorder	1,961,027	20.4	603,400	16.1
Alcohol use disorder	1,561,302	16.2	553,027	14.7
Drug use	1,654,129	17.2	588,481	15.7
Post-traumatic stress disorder	845,842	8.8	400,959	10.7
Eating disorder	350,320	3.6	144,204	3.8
Traumatic brain injury	312,516	3.3	140,940	3.8
Prescription fills in past 5 years				
Antidepressants	6,521,196	67.9	2,559,147	68.2
Benzodiazepines (anxiety/sedative)	4,536,960	47.2	1,641,900	43.7
Hypnotics	1,427,849	14.9	381,353	10.2
2nd generation antipsychotics	2,035,084	21.2	799,401	21.3
Encounters in prior 5 years with mental health diagnosis				
Inpatient	2,296,579	23.9	859,640	22.9
Outpatient	8,789,642	91.5	3,431,776	91.4
ED	3,167,119	33.0	1,347,840	35.9
Recorded suicide attempt in prior 5 years	381,591	4.0	172,329	4.6
Charlson comorbidity index				
0	7,133,017	74.2	2,719,729	72.4
1	1,402,435	14.6	565,038	15.1
>1	1,074,866	11.2	469,370	12.5
Suicide death within 90 days of visit	2318	24 per 100,000	710	19 per 100,000

[1] *The PHQ-9 9th item asks patients about the frequency of thoughts of suicide or death in the prior 2 weeks.*

Table 2.1 shows the out-of-sample AUCs from 5-fold CV for each combination of tuning parameters for the split sample method. Based on the results, the parameter combination with *ntry* = $0.5\sqrt{p}$ and minimum node size = 25,000 had the optimal AUC, which was 0.811.

Table 2.1 Tuning parameter results for split sample estimation

Min node size[1]	*Ntry*[2] = $0.5\sqrt{p}$ [3]	$\sqrt{p}$	$2\sqrt{p}$
10,000	0.800	0.800	0.800
25,000	0.811[4]	0.806	0.808
50,000	0.810	0.803	0.806
100,000	0.807	0.807	0.807
150,000	0.802	0.807	0.805

1 *The minimum node size in the model*
2 *The number of parameters randomly selected for consideration for splitting at each node*
3 *p denotes the total number of predictors*
4 *The optimal AUC among tuning parameter combinations considered*

Table 2.2 shows the out-of-sample AUCs from 5-fold CV for each combination of tuning parameters for the entire sample method. Based on the results, the parameter combination with *ntry* = $\sqrt{p}$ and minimum node size = 100,000 had the optimal AUC, which was 0.832.

Table 2.2 Tuning parameter results for entire sample estimation

Min node size[1]	*Ntry*[2] = $0.5\sqrt{p}$ [3]	$\sqrt{p}$	$2\sqrt{p}$
50,000	0.829	0.829	0.824
100,000	0.827	0.832[4]	0.825
250,000	0.825	0.827	0.826
500,000	0.821	0.822	0.823

1 *The minimum node size in the model*
2 *The number of parameters randomly selected for consideration for splitting at each node*
3 *p denotes the number of total number of predictors*
4 *The optimal AUC among tuning parameter combinations considered*

Table 3 shows the AUCs of prediction models developed using the split sample and entire sample approach in the development and prospective validation datasets. In the split sample method, the AUC of the prediction model was 0.837 (95% CI: 0.826 – 0.848) in the testing set of the development dataset, which was similar to the AUC estimated by 5-fold CV for the prediction model fit to the entire sample (0.832, 95% CI: 0.824 – 0.838). Approximate and exact optimism correction methods for the prediction model fit in the entire sample had higher estimated AUCs (0.920 with 95% CI: 0.917 – 0.924 and 0.911 with 95% CI 0.904 – 0.917 respectively).

Prediction models estimated with the split sample and entire sample approaches had similar AUCs in the prospective validation dataset: the estimated AUC was 0.809 (95% CI: 0.795, 0.822) for the split sample method and 0.811 (95% CI: 0.795, 0.824) for the entire sample method. Comparing prospective performance to the AUC estimates from the development dataset, the split sample and 5-fold CV method both slightly overestimated the AUC in the prospective validation dataset. However, overestimation was much worse for the approximate and exact optimism correction methods; these two methods did not prevent overestimation of the prediction model's future performance using the entire development sample.

Table 3: AUCs (and 95% CIs) of prediction models from split-sample and entire sample estimation approaches on the development dataset and prospective validation dataset

	Split sample	Entire Sample			
		Apparent In-sample[1]	Approximate Optimism Correction	Exact Optimism Correction	5-fold CV
Development dataset	0.837 (0.826, 0.848)	0.926 (0.921, 0.929)	0.920 (0.917, 0.924)	0.911 (0.904, 0.917)	0.832 (0.824, 0.838)
Prospective validation dataset	0.809 (0.795, 0.822)	0.811 (0.795, 0.824)			

[1] *The apparent AUC of the prediction model estimated in the entire sample*

Table 4.1 reports the sensitivity at various risk thresholds for both sampling approaches in the development and prospective validation dataset. For the split sample method, among visits with suicide risk in the highest percentile, estimated sensitivity in the development dataset was 10.95%. This indicates that using this prediction model to select visits in the top 1% of estimated suicide risk correctly identifies 10.95% of visits with suicide death, with a 95% CI (9.38%, 12.93%).

At all thresholds, approximate and exact optimism methods had higher estimates of sensitivity in the development dataset. For example, for a risk threshold defined at the 99th percentile, the estimated sensitivity was 10.95% (9.38%, 12.93%) for the split sample approach and 11.65% (95% CI: 10.43% - 12.62%) for 5-fold CV with the entire sample prediction model, but was much high for the approximate and exact optimism methods, estimated to be 31.96% (95% CI: 28.25% - 34.95%) and 28.69% (95% CI: 26.17% - 31.21%), respectively. Prediction models estimated using the split and entire sample methods had similar sensitivity on the prospective validation dataset: sensitivity of a risk threshold defined at the 99th percentile was 12.26% (95% CI: 9.86% - 14.44%) for the split sample prediction model compared to 15.07% (95% CI: 12.54% - 17.43%) for the entire sample prediction model. More importantly, at most thresholds, both exact and approximate optimism correction methods for the entire sample method overestimated the sensitivity in the prospective validation dataset using the development dataset, though the error was smaller for the exact optimism correction method compared to the approximate method. Overall, the estimated sensitivity from the split sample method and the entire sample method with 5-fold CV better reflected prediction model performance in the prospective validation dataset.

Table 4.1 – Sensitivity (95% CI) in the development and prospective validation datasets of prediction models estimated with a split sample and entire sample approach

Risk Score Percentile Cut-Points	Split Sample		Entire Sample			
	Development dataset (95% CI)	*Prospective validation dataset* (95% CI)	Approximate Optimism (95% CI)	Exact Optimism (95% CI)	5-folds CV (95% CI)	*Prospective validation dataset* (95% CI)
> 99%	10.95% (9.38%, 12.93%)	12.26% (9.86%, 14.44%)	31.96% (28.25%, 34.95%)	28.69% (26.17%, 31.21%)	11.65% (10.43%, 12.62%)	15.07% (12.54%,17.43%)
> 95%	35.97% (33.23%, 38.53%)	41.41% (37.92%, 44.76%)	61.23% (58.95%, 63.25%)	58.58% (55.93%, 61.24%)	33.91% (31.99%, 35.61%)	39.16% (35.78%, 42.25%)
>90%	55.2% (52.25%, 58.59%)	53.10% (49.47%, 56.48%)	75.56% (73.85%, 77.52%)	72.67% (70.17%, 75.16%)	50.47% (48.83%, 52.5%)	52.96% (49.05%, 56.34%)
> 75%	76.17% (73.64%, 78.54%)	73.66% (70.31% ,77.76%)	91.33% (90.08%, 92.33%)	89.74% (88.39%, 91.09%)	75.11% (73.12%,76.45%)	72.11% (68.62%,75.43%)

Table 4.2 reports the specificity at different risk thresholds for each sampling method in the development and prospective validation datasets. Both prediction models estimated using the split sample and entire sample estimation had similar specificity in the prospective validation dataset. For e.g., at the 99th percentile threshold, specificity was estimated to be 98.72% (98.71% - 98.73%) for the split sample method compared to 98.80% (98.79% - 98.81%) for the entire sample method. In addition, split sample estimation, approximate optimism, exact optimism and 5-fold CV produced similar estimates for specificity in the development dataset at all risk thresholds.

Table 4.3 shows the estimated positive predictive value (PPV) at different risk thresholds for each sampling method in the development and prospective validation datasets. The exact and approximate optimism method had the highest estimates of PPV in the development dataset. The estimates of PPV were similar for the split sample estimation and 5-fold CV on the development dataset. For example, for visits with predicted risk above the 95th percentile threshold, the PPV was 164 deaths per 100,000 visits (95% CI: 152 – 176) for the split sample method and 160 deaths per 100,000 visits (95% CI: 151 – 169) for 5-fold CV with the entire sample. In addition, prediction models estimated with the split sample and entire sample approach had similar PPV in the prospective validation dataset. All methods overestimated PPV in the prospective validation set, which was expected because the overall event rate was lower in the prospective validation set than in the development dataset (19 vs. 24 deaths per 100,000 visits, respectively). Similar to other performance measures, both approximate optimism and exact optimism estimates of PPV in the development dataset overestimated observed PPV in the prospective validation set more than the split sample approach and 5-fold CV with the entire sample.

Table 4.2 – Specificity (95% CI) in the development and prospective validation datasets of prediction models estimated with a split sample and entire sample approach

Risk Score Percentile Cut-Points	Split Sample		Entire Sample			
	Development dataset (95% CI)	*Prospective validation dataset* (95% CI)	Approximate Optimism (95% CI)	Exact Optimism (95% CI)	5-folds CI (95% CI)	*Prospective validation dataset* (95% CI)
> 99%	99.03% (99.02%, 99.04%)	98.72% (98.71%, 98.73%)	99.01% (99.00%, 99.01%)	99.01% (99.00%, 99.02%)	99.07% (99.06%, 99.07%)	98.80% (98.79%, 98.81%)
> 95%	95.06% (95.04%, 95.08%)	94.12% (94.09%, 94.14%)	95.01% (95.00%,95.03%)	95.01% (94.99%, 95.03%)	94.91% (94.89%, 94.92%)	94.68% (94.66%, 94.70%)
> 90%	90.05% (90.02%, 90.08%)	88.31% (88.27%, 88.34%)	90.02% (89.99%, 90.03%)	90.01% (89.99%, 90.04%)	89.86% (89.94%, 89.88%)	88.96% (88.93%, 88.99%)
> 75%	75.03% (74.99%, 75.07%)	71.57% (71.52%, 71.61%)	75.01% (74.99%,75.04%)	75.01% (74.97%, 75.06%)	75.02% (74.99%, 75.05%)	71.93% (71.88%, 71.97%)

Table 4.3 – PPV per 100,000 visits (95% CI) in the development and prospective validation datasets of prediction models estimated with a split sample and entire sample approach

Risk Score Percentile Cut-Points	Split Sample		Entire Sample			
	Development dataset (95% CI)	*Prospective validation dataset* (95% CI)	Approximate Optimism (95% CI)	Exact Optimism (95% CI)	5-folds CV (95% CI)	*Prospective validation dataset* (95% CI)
> 99%	254 (210, 295)	180 (145, 213)	770 (684, 842)	691 (630, 752)	300 (269, 325)	237 (196, 274)
> 95%	164 (152, 176)	133 (122, 144)	295 (284, 304)	283 (270, 295)	160 (151, 169)	139 (127, 150)
> 90%	125 (119, 132)	86 (80, 91)	182 (178, 184)	175 (169, 181)	120 (116, 125)	91 (84, 96)
> 75%	69 (67, 71)	49 (47, 51)	88 (87, 89)	87 (85, 88)	72 (71, 74)	49 (46, 51)

We also examine negative predictive value (NPV). Because suicide death is a rare event in this dataset, the estimated NPV is around 100% at all risk thresholds. Therefore, meaningful conclusions cannot be drawn from these results. Estimates of NPV are reported in the Appendix IV.

Chapter 4. DISCUSSION

4.1 CONCLUSIONS

This study compares the accuracy of split sample and entire sample methods, including two optimism correction methods and 5-fold CV, for internal validation of a suicide prediction model estimated with random forest. The study showed that the split sample estimation method and 5-fold CV using the entire sample most accurately estimated AUC, sensitivity, specificity and PPV in a prospective validation dataset. In addition, both the exact and approximate optimism correction methods in the entire sample underestimated prediction model optimism and, thus, overestimated model performance in the prospective dataset. Our results stand in contrast to prior research demonstrating the accuracy of optimism correction methods with logistic regression models. These optimism correction methods inaccurately estimated optimism for a prediction model estimated with random forest in a large dataset with rare events.

Since the entire sample method uses more data for prediction model estimation, we hypothesized that it may better predict outcomes in the prospective validation dataset. However, we saw that split sample and entire sample estimation methods had similar AUC, sensitivity, specificity, PPV and NPV in the prospective validation dataset. Also, we had concerns about the accuracy of entire sample methods with optimism correction due to the potential for overfitting with a random forest model in a large dataset with rare events. The results showed that this concern was warranted, as optimism correction methods underestimated model optimism.

In addition, we found that the split sample prediction model estimation and 5-fold CV with entire sample model estimation provided accurate optimism correction. The split sample approach is preferable for our application; since the entire sample method required a

computationally intensive bootstrap and model reconstruction process, it requires more time and computational power than the split sample method.

4.2 LIMITATIONS AND FUTURE WORK

There are several limitations in this project. To begin with, we only examined random forest prediction models. Due to the flexibility of the random forest model, e.g. exploring unspecified interactions, it has a greater potential for overfitting compared to the logistic regression models used in prior demonstrations of optimism correction methods with logistic regression and a smaller number of predictors. It is possible that using logistic regression with LASSO variable selection or other machine learning methods like artificial neural networks may show different results.

Overall, we found that the entire sample method with optimism correction did not accurately estimate model performance for a random forest prediction model estimated with a large suicide dataset and rare event. However, we cannot conclude that these optimism correction methods are always inaccurate for a broad class of problems, e.g. random forest models or large datasets. Instead, we recommend being cautious if using the optimism correction method for similar prediction modeling settings and comparing performance estimates to those obtained from more reliable methods, such as a split sample approach or using 5-fold CV with entire sample estimation.

REFERENCES

[1] Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., & Moons, K. (2003). External validation is necessary in prediction research. *Journal of Clinical Epidemiology*, 56(9), 826-832. doi:10.1016/s0895-4356(03)00207-5

[2] Charlson M, Szatrowski T. P., Peterson J, Gold J. Validation of a combined comorbidity index. Journal of clinical epidemiology. 1994 Nov 1;47(11):1245-51.

[3]Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association.* 1983;78(382):316-331

[4] Mitchell, M. S., Gude, J. A., Ausband, D. E., Sime, C. A., Bangs, E. E., Jimenez, M. D., Smith, D. W. (2010). Temporal validation of an estimator for successful breeding pairs of wolves Canis lupus in the U.S. northern Rocky Mountains. *Wildlife Biology*, 16(1), 101-106. doi:10.2981/08-068.

[5] Rahman, M. Shafiqur. A comparison of internal validation methods for validating predictive models for binary data with rare event. *Journal of Statistical Research*. 2017, Vol.51, No.2, pp. 131-144.

[6] Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., Shortreed, S. M. (2018). Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. *American Journal of Psychiatry*, 175(10), 951-960. doi:10.1176/appi.ajp.2018.17101167.

[7] Steyerberg, E. W. (2009). *Clinical Prediction Models A Practical Approach to Development, Validation, and Updating*. Cham: Springer International Publishing.

[8] Steyerberg, E. W., & Harrell, F. E., Jr (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, *69*, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005.

[9] Steyerberg, E. W., Harrell, F. E., Jr, Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., & Habbema, J. D. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, *54*(8), 774–781. https://doi.org/10.1016/s0895-4356(01)00341-9

APPENDIX I: LIST OF PREDICTORS IN THE DATASET

Variable Name	Details
Age	Age in years at index visit
First_visit	1/0. First recorded visit for individual
Days_since_prev	Days since previous visit
Female	Gender=Female
dep_dx_pre5y_noi_cumulative	Depression diagnosis (DX), (0,5] years prior to index visit
dep_dx_pre5y	Depression DX, [1,5] years prior
anx_dx_pre5y_noi_cumulative	Anxiety DX, (0,5] years prior
anx_dx_pre5y	Anxiety DX, [1,5] years prior
bip_dx_pre5y_noi_cumulative	Bipolar depression DX, (0,5] years prior
bip_dx_pre5y	Bipolar depression DX, [1,5] years prior
sch_dx_pre5y_noi_cumulative	Schizophrenia DX, (0,5] years prior
sch_dx_pre5y	Schizophrenia DX, [1,5] years prior
oth_dx_pre5y_noi_cumulative	Other Psychosis DX, (0,5] years prior
oth_dx_pre5y	Other Psychosis DX, [1,5] years prior
dem_dx_pre5y_noi_cumulative	Dementia DX, (0,5] years prior
dem_dx_pre5y	Dementia DX, [1,5] years prior
add_dx_pre5y_noi_cumulative	Attention Deficit Disorder (ADD) DX, (0,5] years prior
add_dx_pre5y	ADD DX, [1,5] years prior
asd_dx_pre5y_noi_cumulative	ASD DX, (0,5] years prior
asd_dx_pre5y	ADD DX, [1,5] years prior
per_dx_pre5y_noi_cumulative	Personality disorder DX, (0,5] years prior
per_dx_pre5y	Personality disorder DX, [1,5] years prior
alc_dx_pre5y_noi_cumulative	Alcohol use disorder DX, (0,5] years prior
alc_dx_pre5y	Alcohol use disorder DX, [1,5] years prior
pts_dx_pre5y_noi_cumulative	Post-traumatic stress disorder (PTSD) DX, (0,5] years prior
pts_dx_pre5y	PTSD DX, [1,5] years prior
eat_dx_pre5y_noi_cumulative	Eating disorder DX, (0,5] years prior
eat_dx_pre5y	Eating disorder DX, [1,5] years prior
tbi_dx_pre5y_noi_cumulative	Traumatic brain injury (TBI) DX, (0,5] years prior
tbi_dx_pre5y	TBI DX, [1,5] years prior
dru_dx_pre5y_noi_cumulative	Drug use disorder DX, (0,5] years prior
dru_dx_pre5y	Drug use disorder DX, [1,5] years prior
antidep_rx_pre3m	Antidepressant prescription (rx) (0,3] months
antidep_rx_pre1y_cumulative	Antidepressant rx (0, 1] year
antidep_rx_pre5y_cumulative	Antidepressant rx (0, 5] years
benzo_rx_pre3m	Benzodiazepine rx (0,3] months
benzo_rx_pre1y_cumulative	Benzodiazepine rx (0,1] year
benzo_rx_pre5y_cumulative	Benzodiazepine rx (0,5] years

hypno_rx_pre3m	Hypnotic rx (0,3] months
hypno_rx_pre1y_cumulative	Hypnotic rx (0,1] year
hypno_rx_pre5y_cumulative	Hypnotic rx (0,5] years
sga_rx_pre3m	Second Generation Antipsychotic (SGA) rx (0,3] months
sga_rx_pre1y_cumulative	SGA rx (0,1] year
sga_rx_pre5y_cumulative	SGA rx (0,5] years
mh_ip_pre3m	Inpatient hospitalization with mental health diagnosis (IP with MH dx) (0, 3] months
mh_ip_pre1y_cumulative	IP with MH dx (0, 1] year
mh_ip_pre5y_cumulative	IP with MH dx (0, 5] years
mh_op_pre3m	IP with MH dx (0, 3] months
mh_op_pre1y_cumulative	IP with MH dx (0, 1] year
mh_op_pre5y_cumulative	IP with MH dx (0, 5] years
mh_ed_pre3m	Emergency Department (ED) utilization with MH dx (0, 3] months
mh_ed_pre1y_cumulative	ED with MH dx (0, 1] year
mh_ed_pre5y_cumulative	ED with MH dx (0, 5] years
any_sui_att_pre3m	Prior suicide attempt (0, 3] months
any_sui_att_pre1y_cumulative	Prior suicide attempt (0, 1] year
any_sui_att_pre5y_cumulative	Prior suicide attempt (0, 5] years
lvi_sui_att_pre3m	Laceration-based violent (LVI) suicide attempt (0, 3] months
lvi_sui_att_pre1y_cumulative	LVI suicide attempt (0, 1] year
lvi_sui_att_pre5y_cumulative	LVI suicide attempt (0, 5] years
ovi_sui_att_pre3m	Other (not laceration-based) violent (OVI) suicide attempt (0, 3] months
ovi_sui_att_pre1y_cumulative	OVI suicide attempt (0, 1] year
ovi_sui_att_pre5y_cumulative	OVI suicide attempt (0, 5] years
any_inj_poi_pre3m	Any prior injury or poisoning (0, 3] months
any_inj_poi_pre1y_cumulative	Any prior injury or poisoning (0, 1] year
any_inj_poi_pre5y_cumulative	Any prior injury or poisoning (0, 5] years
any_sui_att_pre5y_cumulative_f	Prior suicide attempt (0, 5] years * (gender=F)
any_sui_att_pre5y_cumulative_a	Prior suicide attempt (0, 5] years * (age)
charlson_score	Charlson comorbidity score on index date
charlson_a	Charlson score on index date * (age)
charlson_mi	Charlson indicator myocardial infaction
charlson_chd	Charlson indicator coronary heart disease
charlson_pvd	Charlson indicator peripheral vascular disease
charlson_cvd	Charlson indicator cerebrovascular disease
charlson_dem	Charlson indicator dementia
charlson_cpd	Charlson indicator chronic obstructive pulmonary disease
charlson_rhd	Charlson indicator rheumatic heart disease
charlson_pud	Charlson indicator peptic ulcer disease

charlson_mlivd	Charlson indicator mild liver disease
charlson_diab	Charlson indicator diabetes
charlson_diabc	Charlson indicator diabetes with complications
charlson_plegia	Charlson indicator paralysis (paraplegia)
charlson_ren	Charlson indicator renal disease
charlson_malign	Charlson indicator malignancy
charlson_slivd	Charlson indicator severe liver disease
charlson_mst	Charlson indicator Metastatic Cancer
charlson_aids	Charlson indicator AIDS
hispanic	Hispanic ethnicity
census_missing	Census variables observed (0/1)
hhld_inc_lt40k	Neighborhood income 40 (<$40k/>=$40k)
coll_deg_lt25p	Neighborhood education (<25% college grad/>=25% college grad)
phqnumber90	Total # of Patient Health Questionnaires 9th item (PHQ9s) in prior 90 days
phqmode90_0	Modal PHQ9 value in prior 90 days = 0
phqmode90_1	Modal PHQ9 value in prior 90 days = 1
phqmode90_2	Modal PHQ9 value in prior 90 days = 2
phqmax90_0	Maximum PHQ9 value in prior 90 days = 0
phqmax90_1	Maximum PHQ9 value in prior 90 days = 1
phqmax90_2	Maximum PHQ9 value in prior 90 days = 2
phqmax90_3	Maximum PHQ9 value in prior 90 days = 3
phqnumber183	Total # of Patient Health Questionnaires (PHQs)_ in prior 183 days
phqmode183_0	Modal PHQ value in prior 183 days = 0
phqmode183_1	Modal PHQ value in prior 183 days = 1
phqmode183_2	Modal PHQ value in prior 183 days = 2
phqmax183_0	Maximum PHQ value in prior 183 days = 0
phqmax183_1	Maximum PHQ value in prior 183 days = 1
phqmax183_2	Maximum PHQ value in prior 183 days = 2
phqmax183_3	Maximum PHQ value in prior 183 days = 3
phqnumber365	Total # of PHQs in prior 365 days
phqmode365_0	Modal PHQ value in prior 365 days = 0
phqmode365_1	Modal PHQ value in prior 365 days = 1
phqmode365_2	Modal PHQ value in prior 365 days = 2
phqmax365_0	Maximum PHQ value in prior 365 days = 0
phqmax365_1	Maximum PHQ value in prior 365 days = 1
phqmax365_2	Maximum PHQ value in prior 365 days = 2
phqmax365_3	Maximum PHQ value in prior 365 days = 3
dep_dx_pre5y_cumulative	Depression DX, [0,5] years prior
anx_dx_pre5y_cumulative	Anxiety DX, [0,5] years prior
bip_dx_pre5y_cumulative	Bipolar depression DX, [0,5] years prior
sch_dx_pre5y_cumulative	Schizophrenia DX, [0,5] years prior
oth_dx_pre5y_cumulative	Other Psychosis DX, [0,5] years prior

dem_dx_pre5y_cumulative	Dementia DX, [0,5] years prior
add_dx_pre5y_cumulative	ADD DX, [0,5] years prior
asd_dx_pre5y_cumulative	ASD DX, [0,5] years prior
per_dx_pre5y_cumulative	Personality disorder DX, [0,5] years prior
alc_dx_pre5y_cumulative	Alcohol use disorder DX, [0,5] years prior
dru_dx_pre5y_cumulative	Drug use disorder DX, [0,5] years prior
pts_dx_pre5y_cumulative	PTSD DX, [0,5] years prior
eat_dx_pre5y_cumulative	Eating disorder DX, [0,5] years prior
tbi_dx_pre5y_cumulative	TBI DX, [0,5] years prior
phq8_index_score_calc	PHQ 1st-8th items (PHQ8) total score at index visit
phq8_missing	PHQ8 missing at index visit (1 if missing)
race1	Race
inscat	Insurance type
phq9	PHQ-9 9th item response at index visit
unenrolled	individual not enrolled in health system on date of visit

APPENDIX II: VERSION FOR R PACKAGES

R (3.6.2)

Tidyverse (1.3.0)

doParallel (1.0.15)

bit64 (0.9.7)

ranger (0.12.1)

sqldf (0.4.11)

data.table (1.12.8)

cvAUC(1.1.0)

foreach (1.4.7)

APPENDIX III: SAMPLE CODE

```
#function that performs a stratified bootstrap and saves information so that it can be used to tell ranger bootstrap samples each tree should be built on
sboot <- function(data, outcome){
  indi <- rep(0,nrow(data))
  ind1 <- data[get(outcome)==1, which=TRUE]
  ind0 <- data[get(outcome)==0, which=TRUE]
  booti <- data.table(ind=c(sample(ind1, replace=TRUE), sample(ind0, replace=TRUE)))
  booti <- booti[, .(freq=.N), by=ind] # make a frequency summary table
  indi[booti$ind] <- booti$freq
  return(indi) ## frequency of each observation to be chosen
}

# 5-fold CV for random forest
auc_cv <- foreach(j=min.node.size, .combine=cbind) %do% {   #code cycles through all options for min.node.size
  auc_nv <- foreach(i=node.vars, .combine=cbind) %do% {    #code cycles through all options for all node.vars
    preds <- foreach(k=1:5, .combine=c) %do% {  #code cylces through CV indices 1-5
      #this next function estimates a random forest model with the selected min.node.size, node.vars. and excluded CV fold
      rf_mod <- ranger(dependent.variable.name="death90",
                       data = training[cvvar != k, c(feat,"death90"), with=FALSE],
                       num.trees = cur.num.trees,
                       mtry=i,
                       importance="none”,
                       write.forest=TRUE,
                       probability=TRUE,
                       min.node.size=j,
                       respect.unordered.factors="partition
                       oob.error=FALSE,
                       save.memory=FALSE,
                       inbag=cv.list[[k]])
      print(c(i,j, k)) #trace for progress
      pre_cv = predict(rf_mod,training[cvvar == k, c(feat), with=FALSE])
      pre_cv$predictions[,2]
       } #end preds
    AUC (preds, labels = training %>% arrange(cvvar) %>% select(death90) )
  } #end auc_nv
  auc_nv
} #end outer foreach

#Function of sensitivity, specificity, PPV and NPV
test_perf = function(cutoff, #cutoffs defined in the training
```

```
                outcome, ## testing dataset results
                pre_te ## predicted risk score for the testing
){
  ns <- TP <- TN <- FP <- FN <- vector(length=length(cutoff))
  for(j in 1:length(cutoff)){
    ns[j] <- sum(pre_te >= cutoff[j]) #number of observations with risk score above the threshold
    TP[j] <- sum(pre_te >= cutoff[j] & outcome == 1) #number of true positives, risk score is
above threshold and person had an event
    FP[j] <- sum(pre_te >= cutoff[j] & outcome == 0) #number of false positives, risk score is
above threshold but person did not have event
    FN[j] <- sum(pre_te < cutoff[j] & outcome == 1) #number of false negatives, risk score is
below threshold but person had event
    TN[j] <- sum(pre_te < cutoff[j] & outcome == 0)  #number of true negatives, risk score is
below threshold and person did not have an event
  }
```

Optimism correction loop

```
while(m < 101 | diffauc > 0.0001 | length(which(diff_other > 0.0001) ) != 0){
  sboot1 = sboot(data_s,'death90')
  sample1 = data.table(data_s[rep(1:nrow(data_s),sboot1)])
```

##calculate the modelo on the bootstrap data

```
  pre_original = sample1$preds_final
  auc_orignal[m] = AUC(pre_original,labels = sample1$death90)
  all_per_original[[m]]= test_perf(ps,sample1$death90,pre_final,pre_original)
```

##predictions using the bootstrap sample

```
  cv.list1 <- vector(mode = "list", length = cur.num.trees)
  for(ii in 1:length(cv.list1)){
    cv.list1[[ii]] <- sboot(data_s, "death90")
  }

  rf_final1 <- ranger(dependent.variable.name="death90",
                data = sample1[, c(feat,"death90"), with=FALSE],
                num.trees = cur.num.trees,
                mtry= ntry,
                importance="none",
                write.forest=TRUE,
                probability=TRUE,
                min.node.size= minnode,
                respect.unordered.factors="partition,
                oob.error=FALSE,
                save.memory=FALSE,
                inbag=cv.list1
  )
```

##find the AUC on the bootstrap set

```
 pre_final1 = predict(rf_final1, data = sample1[,c(feat,'death90'), with =FALSE])$predictions[,2]
 auc_final1<-  AUC( pre_final1,labels = sample1$death90)
 auc_samplem[m] = auc_final1
 ##find the AUC on the original data
 pre_final1o= predict(rf_final1, data = data_s[,c(feat,'death90'),with=FALSE])$predictions[,2]
 auc_final1o<-  AUC( pre_final1o,labels = data_s$death90)
 auc_orignalbt[m] = auc_final1o
 ##calculate optimism for auc
 optimism_auc = auc_final1 - auc_final1o

 ## calculate other test statisticcs
 test_p1 = test_perf(ps,sample1$death90,pre_final1,pre_final1) %>%
   unlist%>%matrix(nrow = 4,byrow = F)
 all_per_samplem[[m]] = test_p1
 ## for the original sample
 test_p1o = test_perf(ps,data_s$death90,pre_final1,pre_final1o)%>%
   unlist%>%matrix(nrow = 4,byrow = F)
 all_per_originalbt[[m]] = test_p1o
 ## for the optimism calculation
 optimism_other = test_p1 - test_p1o
 ##save the optimism value
 optimism_other_boot[[m]] = optimism_other

 #calcaulte previous auc mean
 auc_pre = mean(optimism_auc_boot)
 #save all AUC values in optimism_auc_boot
 optimism_auc_boot[m] = optimism_auc
diffauc = abs(mean(optimism_auc_boot) - auc_pre)

 # storing each test performance seperately
 mean_pre = sapply(optimism_other_boot2, mean)
 for(j in 1:16) {optimism_other_boot2[[j]][m] = optimism_other_boot[[m]][j]}
 ##calculate the difference of optimism for others
 diff_other = abs(sapply(optimism_other_boot2, mean) - mean_pre)
 print(paste0('End of optimism ',m))
 rm(rf_final1) ;gc()


 #---------------for 5 folds cv
 preds_best1 = sample1$preds_5fold
 auc_cv_best1[m] = AUC(preds_best1, labels = sample1$death90)
 test_perf_cv[[m]] = test_perf(ps,sample1$death90,pre_final,preds_best1)
 rm(preds_best1)
 print(paste0('End of 5 folds cv ',m))
 m = m + 1
}
```

APPENDIX IV

NPV per 100,000 visits (95% CI) in the development and prospective validation datasets of prediction models estimated with a split sample and entire sample approach

Risk Score Percentile Cut-Points	Split Sample		Entire Sample			
	Development dataset (95% CI)	*Prospective validation dataset* (95% CI)	Approximate Optimism (95% CI)	Exact Optimism (95% CI)	5-folds CV (95% CI)	*Prospective validation dataset* (95% CI)
> 99%	99,980 (99,979, 99,980)	99,983 (99,983, 99,984)	99,983 (99,983, 99,984)	99,983 (99,982, 99,984)	99,978 (99,978, 99,979)	99,983 (99,983, 99,984)
> 95%	99,985 (99,984, 99,985)	99,988 (99,988, 99,989)	99,990 (99,990, 99,991)	99,989 (99,989, 99,990)	99,983 (99,983, 99,984)	99,988 (99,987, 99,988)
> 90%	99,989 (99,988, 99,989)	99,990 (99,989, 99,991)	99,993 (99,993, 99,994)	99,993 (99,992, 99,993)	99,987 (99,986, 99,987)	99,990 (99,989, 99,991)
> 75%	99,993 (99,992, 99,994)	99,993 (99,992, 99,994)	99,997 (99,997, 99,998)	99,997 (99,996, 99,997)	99,992 (99,991, 99,992)	99,993 (99,992, 99,994)