Evaluation of Prediction Performance Metrics in the Rare Event Setting

Emily Minus

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2023

Committee:

Yates Coley

Brian Williamson

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Evaluation of Prediction Performance Metrics in the Rare Event Setting

Emily Minus

Chair of the Supervisory Committee:
Yates Coley
Department of Biostatistics

Area under the receiving operator characteristic curve (AUC) is a commonly reported measure of discriminative performance for binary prediction models. However, there are concerns about AUC being a misleading measure of prediction performance in the rare event setting. This setting is commonly encountered with clinical prediction models, since many events of clinical importance, such as suicide, occur only rarely. We conducted a simulation study to investigate what drives inaccurate or unstable AUC performance in the rare event setting. Specifically, we aimed to determine whether a small number of events is the main driver of the poor AUC performance, or if the main driver is truly the event rate (i.e., there are many events, but they represent a small fraction of the total observations). We also investigated the behavior of other commonly used measures of prediction performance, such as PPV, accuracy, sensitivity, and specificity. Our results indicate that poor AUC performance—as measured by empirical bias, empirical MSE, variability of cross-validated AUC estimates, and empirical coverage of bootstrap intervals—is driven by the number of events, not event rate. While which measure of model performance is of greatest interest depends on how a model will be used, AUC is reliable in the rare event setting provided that the total number of events is moderately large.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Yates and Brian, thank you so much for all your support and guidance during the *entire* process of writing this thesis.

DEDICATION

to my grandfather, who always encouraged my love of math and science

Chapter 1

INTRODUCTION

1.1 Clinical prediction models

Predictive models for use in the clinical setting are of interest as a way to obtain more accurate prognoses and diagnoses [38], guide clinical decision-making through early warning systems [39], identify patients at risk for an adverse outcome [44], and better allocate health resources [9]. Some traditional severity scales, such as APACHE II in the critical care context, are constructed by scoring each item and then summing to obtain the overall severity score. In APACHE II, for example, each item is a physiological measure and is scored based on the deviation from "normal" values [27]. Tools such as logistic regression, penalized regression, or tree-based methods offer a different approach to creating predictive models, and these tools can also provide a continuous risk score. In some contexts, this continuous risk score is of interest. However, in the clinical setting predictive models can also be useful when they serve as screening tests to identify patients at risk [9]. In these cases, where the outcome of interest is binary, a continuous (0-1) risk score output by a prediction model is interpreted as the predicted probability of experiencing the outcome.

To use a predictive model that outputs a continuous risk score as a binary classifier, there has to be a *decision threshold*. Anything with a risk score higher than this threshold is a predicted event, while anything lower than this threshold is a predicted non-event. The predicted event statuses can fall into four possible categories: false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN); these four quantities make up the *confusion matrix*. If we think of characterizing performance with proportions or rates instead of counts, we can show that the true positive rate (TPR, also known as sensitivity) and false positive rate (FPR, 1 - specificity) together are sufficient to completely characterize both the information from the confusion matrix and the classifier performance [35]. Various measures of prediction performance can be calculated from the confusion matrix,

and different prediction performance metrics provide different information. Which metric is of greatest importance depends on the context in which a predictive model will be used, and looking at multiple metrics can give a fuller understanding of classifier performance. However, all measures of prediction performance calculated from the confusion matrix are dependent on the decision threshold, because the relative numbers of false positives, false negatives, true positives, and true negatives—and correspondingly the false positive rate and and true positive rate—depend on the decision threshold.

1.2 The rare event setting

Often, clinically important events occur only rarely. An example of a rare, clinically important event is suicide. In one paper presenting a model predicting suicide death and suicide attempt, a sample of 2.96 million patients with 10,275,853 mental health visits and 9,685,206 primary care visits had a per-visit suicide attempt rate of 0.62% and 0.26% for mental health visits and primary care visits, respectively. Suicide deaths were even more rare, with a per-visit event rate of 0.02% for mental health visits and 0.01% for primary care visits [44]. Despite this rarity, suicide was the 10th leading cause of death in 2019 (the last year before the SARS-CoV-2 pandemic). In 2019 there were over $47,000$ suicide deaths, representing 13.9 suicide deaths per $100,000$ population [28]. Predicting suicide deaths and attempts in order to identify individuals at risk—so that these individuals can be offered services—is clearly of interest. A 2019 systematic review found a total of 64 unique suicide prediction models [4]. The rare event setting, however, presents some challenges to prediction. For one, even with a large number of observations the absolute number of events may be small. In predictive modeling of binary outcomes, the number of events, not just the total sample size, is of importance. Additionally, a small event rate impacts the interpretation of some common measures of predictive model performance, because the overall event rate (specifically, balanced versus unbalanced data) impacts the behaviors of these measures [32, 43].

1.3 What is AUC?

Setting aside the rare event setting for a moment, the dependence on a decision threshold also complicates the use of measures such as false positive rate and true positive rate to compare or describe classifier performance. One alternative to choosing a single decision threshold is to plot the *FPR* (x-axis) and *TPR* (y-axis) against each other as the decision threshold varies over all possible values. The *FPR* and *TPR* must vary with each other as the decision threshold changes, and the result is a curve called the Receiver Operating Characteristic (ROC) curve. An ROC curve that lies more toward the upper left-hand corner of the ROC space indicates better performance than a curve that lies closer to the $x = y$ diagonal [35].

While ROC curves can be compared qualitatively, a one-number summary of classifier performance is desirable for model comparison [46]. The area under the ROC curve (AUC) is such a one-number summary of the ROC curve. The true AUC of a classifier can be interpreted as the probability of correctly ranking a randomly selected event/non-event pair [19]. In the setting of a predictive model that outputs a predicted probability, let X be the predictors, Y be the binary outcome with $Y = 1$ indicating an event and $Y = 0$ indicating a non-event, and $f(X)$ the predicted probability. Then the true AUC is

$$\text{AUC}_{\text{true}} = P[f(X_i) > f(X_j) \mid Y_i = 1, Y_j = 0].$$

Correspondence between this meaning of AUC and the Wilcoxon statistic gives a formula for estimating AUC directly from the data and classifier output [19]. Using the same notation as above and letting n be the number of observations,

$$\text{AUC} = n^{-2} \sum_{i=1}^{n} \sum_{j \neq i} I(f(x_i) > f(x_j) \mid y_i = 1, y_j = 0).$$

AUC is a measure of discriminative performance. Discrimination refers to relative predictive accuracy, i.e., how well a model separates events from non-events. Calibration, in contrast, refers to absolute predictive performance, i.e., how well predicted probabilities align with the observed outcome proportions in a new sample [22]. As implied by the in-

terpretation of the true AUC as a probability, AUC can take values between 1 and 0. An AUC of 1 indicates perfect discrimination. In all possible event/non-event pairs (y_i, y_j) with $y_i = 1$ and $y_j = 0$, $f(x_i) > f(x_j)$. An AUC of 0 also indicates perfect discrimination, but (problematically) with predicted probabilities of events lower than predicted probabilities of non-events: $f(x_i) < f(x_j)$ for all event/non-event pairs. An AUC of 0.5 indicates the worst possible discriminative performance. In 50% of pairs, $f(x_i) > f(x_j)$ and in 50% of pairs $f(x_i) < f(x_j)$, i.e. the classifier performs no better than assigning predicted probabilities at random [7].

1.4 Arguments for and against the use of AUC

There are various arguments for and against the use of AUC in evaluating the performance of predictive models. One argument for AUC is that it satisfies three characteristics of a good measure of the performance of a binary classifier. A good measure of classifier performance should be: 1) a single number with the same scale for all systems; 2) objective, i.e., not reliant on the subjective decisions of the people creating the model; and 3) relatively independent of the event rate [46]. A single number makes comparisons of classifiers straightforward, especially in comparison to graphical measures of performance such as the ROC curve or the precision-recall curve [7, 46]. Unlike accuracy, sensitivity, or specificity (among other performance measures), AUC is also objective. While these other measures require the specification of a decision threshold for classifying predicted probabilities or scores as events or non-events, AUC averages over all possible thresholds [15]. This independence from the choice of threshold is especially desirable when there is no clear rationale for the decision threshold [7]. The decision threshold controls the bias toward false positives and false negatives [46], but only vary rarely can we know with any certainty, or even agree upon, the relative costs of false positives versus false negatives [7]. In fact, the optimal trade-off between false negatives and false positives may vary by setting [15]. Finally, in comparison to performance measures such as accuracy, AUC is also relatively un-influenced by the event rate [7, 22, 46].

Another argument for the use of AUC as a measure of model performance is that it is an interpretable, intuitive metric of discriminative performance. Since AUC is based on the

ranking or ordering of a classifier, specifically pairwise comparisons of predicted probabilities or scores of events and non-events, we can think of it as a "natural criterion" for discrimination [12]. This is reflected in the interpretation of AUC as the probability of correctly ranking an event/non-event pair [19]. In some contexts, this discriminative performance is the most relevant aspect of model performance to capture. In the setting of clinical prediction models, an example of such a context is when the ultimate aim of a prediction model is to identify those at highest risk so that they can be offered additional resources. When those resources are limited, the goal is to correctly discriminate the individuals who would most benefit from those resources from those who would receive no benefit or less benefit [45]. AUC is also intuitive in that it incorporates an understanding that false positive rate and true positive rate are linked and will vary together as the decision threshold changes [46]. Finally, some authors argue for the use of AUC because it performs better in model selection than accuracy. Ling, Huang, and Zhang (2003) showed that AUC is statistically consistent with accuracy, but is more *discriminating* than accuracy.

More discriminating, here, refers to the idea that there will be fewer "ties" between models with AUC than with accuracy [33]. Rosset (2004) demonstrated, on both real and simulated datasets, that in many situations out-of-sample AUC can better select the models that minimize future misclassification error than out-of-sample misclassification error. While out-of-sample AUC might be biased for estimating misclassification error, intuitively it has lower variance than out-of-sample misclassification error because AUC is calculated using more comparisons [42].

Many have also made arguments against the use AUC. Some of these arguments take the view that calibration, not discrimination, is the most important dimension of model performance. This view that calibration is more important is based on the idea that assessing or predicting risk is fundamentally prognostic, not diagnostic, in nature [22, 45]. Since AUC does not measure the calibration of a model, it is then follows that it is impossible to interpret how "good" a model is based on its AUC [16, 22]. A model can have very poor calibration but high AUC, since predicted probabilities do not impact AUC beyond their role in ranking [11]. Furthermore, in real world settings, perfect discrimination (AUC = 1) and perfect calibration are incompatible. Assuming a uniform distribution of true risk,

the maximum AUC obtainable by a model with perfect calibration is only 0.83 [14]. Under more realistic assumptions of risk distribution, specifically cardiovascular risk, Cook (2007) suggested that the maximum AUC of a perfectly calibrated model would be between 0.75 and 0.9 [11]. In addition to this argument against AUC based on the view that calibration is the most important measure of performance, there have been arguments against the use of AUC based on the views that post-test probability (e.g., the probability of having the event given classification into the high-risk group) is most relevant [37], or that AUC may be undesirable because it aggregates over pairs of false positive and true positive rates that would never be of practical significance. This last point was made in discussing potential benefits of using partial-AUC instead of AUC in certain contexts [15].

AUC is also sometimes considered insensitive, in that adding new, important predictors to a model may result in only small increases in AUC. Cook (2007) shows this insensitivity of AUC in the context of models for cardiovascular events and predictors such as smoking history or serum cholesterol levels. She argues that if AUC were to be used as the only metric of model selection, this insensitivity would result in erroneous elimination of candidate predictors that meaningfully improve calibration—specifically, the stratification of patients into risk categories—but that only minimally improve AUC. This argument takes the view that in this context of predicting cardiovascular risk the calibration of a model is of greater importance than its discriminative performance [11].

There are also concerns that AUC can be misleading. If ROC curves for two predictive models cross, their AUCs may be identical but one model may have meaningfully worse performance than the other in the region of decision threshold values that are clinically relevant. Even when the ROC curves do not cross, the AUCs may be similar even though there are meaningful differences in performance in the region of relevant threshold values [34]. There are especially concerns about AUC, and ROC curves more generally, being misleading or uninformative in the rare event setting [2, 6, 32, 43].

1.5 Poor AUC performance in the rare event setting

In a recent paper, Adhikari et al. (2021) suggested that AUC may be problematic in the rare event setting and recommended that in this setting AUC should not be reported as a

primary metric of prediction performance [2]. Using simulated data of varying complexity as well as data from a state-mandated clinical registry of aortic valve replacements, the authors demonstrated that even when AUC is high other measures of model performance can be poor. In particular, they found that high AUC could be accompanied by a very low true positive rate at the threshold chosen to optimize accuracy. Thus, they argued that AUC is misleading in the rare event setting and so should be avoided. Instead, Adhikari et al. recommended reporting a suite of other prediction metrics [2]. In contrast to this recommendation of avoiding AUC in the rare event setting, in their work creating suicide risk prediction models, biostatisticians at the Kaiser Permanente Washington Health Research Institute have found AUC to be a useful metric of prediction performance in this setting and have not observed issues with its use. While anecdotal, this observation motivated this complementary simulation study investigating the reliability of prediction metrics—and what drives that reliability—in the rare event setting.

1.6 Goals of this simulation study

The goal of this simulation study is to investigate what drives inaccurate or unstable AUC performance in the rare event setting. As discussed above, there are many arguments for and against the use of AUC as a measure of predictive model performance. However, in the rare event setting, a commonly encountered sentiment is that AUC is an "incorrect" measure of model performance. By demonstrating the statistical properties of AUC in this setting, our goal was to shift the question from "is AUC correct in the rare event setting" to "is AUC a performance measure of interest" given the context of prediction.

Specifically, we aimed to determine if a small number of events is the true driver of poor AUC performance, or if the true driver is event rate. While the focus of this simulation is AUC, in concordance with the view that evaluating performance using a suite of prediction metrics provides a fuller understanding of model performance, we also investigated the behavior of multiple other metrics. Understanding how the reliability of each of these metrics depends differently on event rate, the number of events, and the number of non-events is useful for interpreting their meaning. Additionally, we investigated if methods such as stratified data splitting, averaging over multiple rounds of cross-validation to obtain

metric estimates, or using a certain type of confidence intervals can guard against poor behavior.

Chapter 2

DATA

Data for each simulation was generated by sampling without replacement from a *full dataset* of 3,081,420 observations derived from an existing Suicide Risk Supplement (SRS) dataset. This SRS dataset contains 25 million outpatient mental health visits by 3 million patients, where an outpatient mental health visit is defined as a visit to a mental health provider or a primary care visit with a mental health diagnosis. Seven sites, all members of the Mental Health Research Network, contributed data to the SRS dataset: HealthPartners; Henry Ford Health System; and the Colorado, Hawaii, Northwest, Southern California, and Washington regions of Kaiser Permanente. The SRS dataset consists of all outpatient mental health visits by patients 13 years of age or older that occurred within these health systems between January 1, 2009 and September 30, 2017. [10]

The full dataset was created by randomly sampling one mental health visit per patient. For individuals with an event at any visit, a visit with an event was sampled. Thus, each observation in the full dataset corresponds to a separate individual. The binary event was suicide death within 90 days of the mental health visit. The true event rate in the full dataset, R_0, was 0.918%. There were 149 predictors. These predictors consisted of demographic characteristics, mental health and substance abuse diagnoses, other medical diagnoses, past suicide or self-injury attempts, past inpatient or emergency mental health care, dispensed medications, and PHQ-8 score [10, 44]. All predictors except patient age in years and PHQ-8 score (ordinal 0-8) were coded as binary indicators.

Chapter 3

METHODS

3.1 Overview of simulations

In this simulation study we varied the size of the training set and the event rate, allowing us to investigate the reliability of prediction metrics in different settings of sample size and event rarity. Specifically with regard to AUC, varying the training set size allowed us to investigate whether a small number of events is the main driver of the poor AUC performance sometimes observed in the rare event setting, or if the main driver is truly the event rate (i.e., there are many events, but they represent a small fraction of the total observations).

In addition to varying the training set size and event rate, we considered both *naive* and *stratified cross-validation sampling strategies* in order to investigate if sampling strategy can help guard against undesirable prediction metric behavior in the rare event setting. Under a naive sampling strategy, we randomly sampled event and non-event observations in aggregate. For division of data into cross-validation folds, this sampling strategy often results in folds with differing number of events and differing event rates. Under a stratified sampling strategy, we randomly sampled event and non-event observations separately in order to obtain equal–or due to non-divisibility close to equal–number of events and event rates across cross-validation folds.

In total, we conducted a total of 84 simulations by varying the training set size, event rate, and sampling strategy and considering three different prediction algorithms: logistic regression, ridge logistic regression, and random forest with probability trees. We considered five different training set sizes (5,000, 10,000, 50,000, 100,000, and 1 million), three different event rates ($R_0/2 = 0.459\%$, $R_0 = 0.918\%$, and $R_0 \times 2 = 1.836\%$), and two different sampling strategies (naive and stratified). For the largest training set size of 1 million observations, we only used event rates $R_0/2$ and R_0 because we had an insufficient number of events in

the full dataset to create a training set and test set both of size 1 million and event rate of $R_0 \times 2$. Otherwise, however, we conducted simulations with all possible combinations of training set size, event rate, sampling strategy, and algorithm. Each simulation consisted of 1000 (for ridge regression and logistic regression) or 500 (for random forest) replications of the procedure. Only 500 replications were done for random forest due to limits on computing time.

Diagrams of the main simulation procedure are provided in Appendix A and outlined in detail in later chapter sections. An overview of the procedure is provided below in Figure 3.1. While AUC was the primary prediction performance metric of interest, we also calculated Brier score, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F_1 score, and $F_{0.5}$ score.

Simulation Procedure Overview	
Inputs: event rate, training set size, algorithm (and hyperparameters), CV sampling strategy	
1	Sample training and test sets from the full dataset
2	Perform 10-fold cross validation to estimate prediction metrics and obtain out-of-sample predicted probabilities
3	Save CV prediction metrics and out-of-sample fold predicted probabilities
4	Bootstrap with out-of-sample fold predicted probabilities to obtain 95% CIs for prediction metrics
5	Fit model on entire training set
6	Predict on test set to obtain test set prediction metrics
7	Save CV prediction metrics, test set prediction metrics, and 95% CIs

Figure 3.1: Overview of the simulation procedure.

3.2 Computing

All simulations and analyses were conducted in R version 4.1.1 [41]. Logistic regression was implemented using the base `R stats` package, penalized regression using the package `glmnet` (version 4.1-4) [17], and random forest using the package `ranger` (version 0.14.1) [48]. The `foreach` package was used to implement parallel computing [13]. For reproducibility, each

replication of the main simulations used a unique random seed from a pre-defined sequence of random seeds.

3.3 Evaluation of prediction metric performance and reliability

For each simulation and for each performance metric, we measured performance by calculating following: the empirical bias of the cross-validated metrics for estimating the test-set performance; the empirical coverage of the 95% confidence intervals for the test-set performance metrics; the empirical mean squared error of cross-validated performance metrics for estimating the test-set performance metrics; the average width of the 95% confidence intervals; and the empirical variability of the cross-validated performance metrics. Table 3.1 provides details on how these measures of performance were calculated from the simulation replicates.

Performance measure	Formula or measure(s)
Empirical bias	$\frac{1}{r}\sum_{i=1}^{r}(m_{i,\text{CV}} - m_{i,\text{test}})$
Empirical coverage of 95% CI	$\frac{1}{r}\sum_{i=1}^{r} I_{95\%\text{CI}}(m_{i,\text{test}})$
Empirical mean squared error	$\frac{1}{r}\sum_{i=1}^{r}\left(m_{i,\text{CV}} - \frac{1}{r}\sum_{j=1}^{r} m_{j,\text{test}}\right)^2$
Empirical variability	variance, range, and IQR of m_{CV}

Table 3.1: Calculation of performance measures from the simulation replicates. Total number of simulations r (500 or 1000), cross-validated performance metric m_{CV}, test-set performance metric m_{test}, and $I_{95\%\text{CI}}(x)$ is an indicator function that equals one if x falls within the bounds of the 95% CI (inclusive on both sides), zero otherwise.

3.4 Data generation

Data for each simulation iteration was generated by sampling without replacement from the full dataset of 3,081,420 observations derived from the Suicide Risk Supplement (SRS) dataset. To create the training set of size n_{train} and test set of size 1×10^6, both with event rate R, we sampled $\lceil R \cdot n_{\text{train}} \rceil$ and $\lceil R \cdot (1 \times 10^6) \rceil$ independent event observations

and $n_{\text{train}} - \lceil R \cdot n_{\text{train}} \rceil$ and $1 \times 10^6 - \lceil R \cdot (1 \times 10^6) \rceil$ independent non-event observations, respectively.

3.5 Algorithms

We considered three algorithms: logistic regression, ridge logistic regression, and random forest. Logistic regression was selected due to its relative stability and its appropriateness for the binary outcome. Ridge regression, selected as an example of a penalized regression method, is a form of penalized regression using the ℓ_2 shrinkage penalty $\lambda \sum_{j=1}^{p} \beta_j^2$, where p is the number of predictors and β_j are the regression coefficients [25]. Random forest is an ensemble tree-based method where a forest of B trees corresponding to B bootstrap samples of the training set is constructed and then predictions obtained by averaging over the trees (or majority voting, in the case of classification trees) in the forest. For each split in each tree, a random sample of m predictors are considered, which de-correlates the trees. Additionally, each tree is grown to a specified depth or minimum node size. [26]. We used probability trees to obtain predicted probabilities of the event occurring. Tree-based methods such as random forest allow for many interactions between predictors without explicitly coding these interactions, in contrast to logistic regression or ridge regression. In these simulations, we did not include interactions between our predictors for logistic regression or ridge regression.

3.5.1 Algorithm hyperparameters

The regularization parameter λ for ridge regression and the minimum node size, number of predictors m considered at each split, the depth of tree, and number of trees B for random forest forest are all hyperparameters. With the exception of the number of trees B, these hyperparameters control the bias-variance trade-off in the algorithms, and are selected to optimize test-set performance [25, 26]. One common approach is to tune over a grid of possible hyperparameter values using cross-validation on the training set. Cross-validation is used to obtain an estimate of test set prediction performance at each grid value, then the grid value at which optimum performance is achieved is selected. Using cross-validation for hyperparameter selection in this way can be computationally intensive.

The goal of this simulation study was not to optimize the prediction performance of a given model but rather to understand how the reliability of measures of prediction performance depend on event rate, training set size, and sampling strategy. For each simulation replication, the model fit on the training set needed to use generally appropriate hyperarameter values, but those hyperparameter values did not need to the values optimal for the replication-specific training set sampled. Given the overall aim of the simulation study, we sought to investigate if we could save computation time by selecting fixed hyperparameters — i.e., for each replication of a simulation with a given rate, training set size, sampling strategy use the same hyperparameter values for the ridge regression or random forest algorithms — instead of performing cross-validation with in the sampled training set to select the hyperparameter(s) in each simulation replication. We determined that we would use fixed hyperparameters if there was stability in the hyperparameter selected across training sets sampled from the full data; if there was stability in the 10-fold cross-validated AUC across different values of the hyperparameter parameter selected; or if there was little difference in the 10-fold cross-validated AUC obtained by using a fixed hyperparameter compared to that obtained by using cross-validation to select the hyperparameter on each replicate-specific sampled training set separately.

For random forest, which has multiple hyperparameters, we only considered tuning over the minimum node size. The number of predictors m randomly sampled as candidates at each split was kept constant at 12, the floor of the square root of the number of features $p = 149$. Previous work has suggested that setting m equal to the square root of the number of features is generally a reasonable default setting [5], and previous work with this SRS data had suggested that this default in reasonable to use here. The maximum number of nodes a tree could have was not restricted beyond the natural restriction imposed by the minimum node size. The number of trees grown, B, was also kept fixed for a given training set size. Evaluation via visual inspection of the increase then plateau in ten-fold cross-validated AUC as number of trees in the forest increased showed that a smaller number of trees could be used for the larger training set sizes, and Table 3.6 shows the number of trees used for each training set size.

3.6 Investigating the use of fixed hyperparameters

Prior to running the main simulations, we conducted an investigation where we used an identical data-generating mechanism and main processes to the full simulations, but with a smaller number of replications, to investigate whether we could use a fixed hyperparameter for each combination of algorithm (random forest, ridge regression), sample size ($n_{\text{train}} = 5,000, 10,000, 50,000, 100,000,$ or 1 million), event rate ($R_0/2, R_0, 2 \times R_0$), and sampling strategy (naive or stratified) instead of selecting the hyperparameter via cross-validation on the sampled training set in each simulation replicate.

This investigation consisted of two parts: 1) nested cross-validation for tuning parameter selection (inner folds) and estimating AUC (outer folds); and 2) a small number of simulations, with relatively few replications, to assess if there was a large difference between the cross-validated AUC when the selected fixed hyperparameters were used and the cross-validated AUC obtained when we used 20-fold cross-validation to select the hyperparameter in the sampled training set. We referred to the second part of this fixed hyperparameter investigations as *pilot simulations*.

3.6.1 Determining and assessing fixed hyperparameters using nested cross-validation

Nested cross-validation with ten outer folds and twenty inner folds was utilized to select the fixed hyperparameters and assess the stability of the cross-validated AUC and selected hyperparameters across cross-validation rounds (Figure A.1, Appendix A). In each replicate of the procedure, a training set of size n_{train} and rate R was randomly sampled from the full dataset. The training set was then divided into ten outer cross-validation folds, using either a stratified sampling strategy or a naive sampling strategy. For each outer hold-out fold, the nine training folds were further divided into twenty inner cross-validation folds, again using either a stratified or naive sampling strategy. The sampling strategy utilized was always consistent across inner and outer folds. Cross-validation for hyperparameter parameter selection was then performed within the inner folds. For both random forest and ridge regression, the value of hyperparameter selected was the value that maximized the 20-fold inner cross-validated AUC, i.e., the average over the 20 folds of the out-of-sample

AUC.

For random forest, tuning was performed over the minimum terminal node size for the probability trees. The grid of minimum node sizes tuned over was a subset of the maximal set of minimum node sizes $\{10, 100, 1000, 10000, 25000, 50000\}$ (Table 3.2). The subset differed by training set sample size such that the minimum node size never exceeded or was equal to the sample size size and, for the larger sample sizes, the smallest minimum node sizes were dropped. For sample sizes greater than 1×10^4, we did not tune over a minimum nodes size of 10. For sample sizes greater than 5×10^4, we additionally did not tune over a minimum node size of 100. We also did not tune over a minimum node size of 1,000 for a training set size of 1×10^5. Finally, we did not tune over a minimum node size of 25000 for the sample size of 5×10^4, since we felt this minimum node size was unlikely to be selected as optimal in this setting.

Training set size	Grid of minimum node sizes
5×10^3	$\{10, 100, 1000\}$
1×10^4	$\{10, 100, 1000\}$
5×10^4	$\{100, 1000, 10000\}$
1×10^5	$\{1000, 10000, 25000, 50000\}$
1×10^6	$\{10000, 25000, 50000\}$

Table 3.2: Grid of minimum node sizes tuned over for each training set size, random forest.

For ridge regression, the package `glmnet` internally computes a sequence of (by default) 100 values of the regularization parameter λ [17]. We tuned over this internally generated sequence of 100 values of λ. Of note, however, is that the sequence differs depending on the training data supplied, so for each outer fold and replication of the procedure the exact λ sequence differed slightly.

Using the value of the hyperparameter selected to minimize the 20-fold inner cross-validated AUC, we then re-fit the algorithm on all the data contained within the outer training folds. Using this model, we predicted on the outer hold-out fold, then used these predictions to calculate the out-of-sample AUC. By repeating this procedure for each of the ten outer folds and averaging over the out-of-sample AUCs from the 10 outer folds, we

obtained the 10-fold outer cross-validated AUC estimate. We also obtained a set of ten selected hyperparameter values, one for each outer cross-validation fold.

For each combination of rate, sampling strategy, and training set size we first completed ten replications of the nested cross-validation procedure. After these ten replications, we calculated the standard deviation of the cross-validated AUCs. If the standard deviation was less than 0.01 (indicating stability of the cross-validated AUC), or if there was stability in the values of hyperparameter selected in the inner rounds of cross-validation, no further replications were done. For ridge regression and the regularization parameter λ, we considered there to be stability in the value of λ selected if, across all outer folds of all replications of nested cross-validation, the standard deviation of the selected λ values was less than 0.01. For random forest, we considered there to be stability in the minimum node size selected if the same minimum node size grid value was selected in all or nearly all outer folds and replications. If the standard deviation of the cross-validated AUC was greater than or equal to 0.01 and there was not clear stability in the values of the selected hyperparameters, another 40 replications were completed. We then calculated the standard deviation of the cross-validated AUCs from the 50 total replications. Again, if the standard deviation of the 10-fold outer cross-validated AUC was less than 0.01 or if there was stability in the selected hyperparameters, no further replications were completed. Otherwise, 50 more replications were conducted, for a total of 100 replications of the nested cross-validation procedure. At the point where the standard deviation of the AUC was less than 0.01 or there was stability in the selected hyperparameters (10 or 50 replications), or after 100 replications, we defined the fixed tuning parameter—for the given combination of training set sample size, event rate, and sampling strategy—as the mode (random forest) or median (ridge regression) of the hyperparameter values selected from the all aggregated rounds of inner cross-validation (Figure A.2, Appendix A).

3.6.2 Pilot simulations for further assessing fixed hyperparameter validity

To further assess the validity of the nested cross-validation procedure for selecting fixed tuning parameters, we used the pilot simulations to check if the 10-fold cross-validated

Decision Tree for Nested Cross-Validation Continuation	
1	Do 10 (40, 50) rounds of nested cross validation
2	Using all rounds, calculate SD(CV-AUC) and SD(λ) or frequencies at which the minimum node sizes in the grid are chosen
3	Assess if SD(CV-AUC) < 0.01, SD(λ) < 0.01), or one minimum node size in grid selected with high frequency
4	If stability in hyperparameter or CV-AUC, or 100 total rounds of nested-cross-validation, go to 5. If there is not stability and less than 100 total rounds of nested cross-validation, repeat 1-3
5	Selected fixed hyperparameter as median(λ) or mode(minimum node size)

Figure 3.2: Decision tree for the rounds of nested cross-validation.

AUC differed meaningfully when using the fixed hyperparameter versus a replication-specific cross-validation-selected hyperparameter. These pilot simulations (Figure A.3, Appendix A) used the same data-generating mechanism and main processes as the nested cross-validation and main simulations, a small number of replications (50), and three combinations of event rate, sampling strategy, and sample size (Table 3.3). For random forest, since after 100 replications of the nested cross-validation procedure with a the combination of $n_{\text{train}} = 5 \times 10^3$, rate $= R_0/2$, and a naive sampling strategy there was a tie for the mode between a minimum node size of 100 and 1,000, we conducted *Pilot Simulation #1* once with a minimum node size of 100 as the fixed hyperparameter and once with a minimum node size of 1000 as the fixed hyperparameter.

We considered these three combinations of training set size, event rate, and sampling strategy (Table 3.3) because the nested cross-validation procedure suggested that there was greater variability in 10-fold cross-validated AUC and hyperparameters selected at a smaller training set sizes, smaller event rates, and a naive sampling strategy. Thus, we believed that if there were to be a large difference in the 10-fold cross-validated AUC when using the fixed hyperparameter compared to a replication-specific 20-fold cross-validation-selected hyperparameter the difference would be largest in this setting. The three combinations of *Pilot Simulations #1, #2,* and *#3* were chosen as representative examples of small training

set size combinations. Using smaller event rates ($R_0/2$ and R_0) and a naive sampling strategy for the smallest training set size and the largest event rate ($R_0 \times 2$) and a stratified sampling strategy for a training set size of 1×10^4 facilitated a comparison of the differences in cross-validated AUC in more or less "difficult" settings.

Pilot Sim.	n_{train}	Rate	Sampling strat.	λ_{fixed}	Fixed min. node size
Sim. # 1	5×10^3	$R_0/2$	naive	0.398191	100 or 1000
Sim. # 2	5×10^3	R_0	naive	0.210478	100
Sim. # 3	1×10^4	$R_0 \times 2$	stratified	0.049776	100

Table 3.3: Training set size, event rate, sampling strategy, along with the fixed regularization parameter and fixed minimum node size(s) used, for the three pilot simulations.

For each replication of the pilot simulation, we sampled a training set of size n_{train} and rate R. This training set was then broken into 20 folds using the specified sampling strategy, and we used 20-fold cross-validation, with same procedure as the inner cross-validation described above, to select the optimal hyperparameter. The hyperparameter selected via this cross-validation was then compared to the relevant fixed hyperparameter. If the hyperparameters were equal, this was recorded and the replication ended. If the hyperparameters differed, 10-fold cross-validation to obtain an estimate of the cross-validated AUC was performed using the hyperparameter selected via 20-fold cross-validation on the same training set and then again using the fixed tuning parameter. We then calculated the difference between these two cross-validated AUC values.

After 50 replications of this procedure, we averaged (over the replications where the cross-validated-selected tuning parameter and the fixed tuning parameters differed; if the tuning parameters were identical, then the CV-AUC would be identical) the difference in the AUC values. This average difference provided an indication as to if using a fixed tuning parameter meaningfully affected the value of cross-validated AUC obtained, in comparison to using cross-validation to select the tuning parameter in each each separate replication of the simulation procedure.

3.6.3 Fixed hyperparameter values used in the main simulation

The stability of the hyperparameters parameters selected in the nested cross-validation procedure (Section 3.6.1), the variability (as measured by standard deviation) of the cross-validated AUC, and the results of the pilot simulation (Section 3.6.2) were all considered in assessing the validity of using fixed turning parameters in the main simulations. The stability of the hyperparameters selected refers, in this case, to how much variation was present in the value of hyperparameter that maximized cross-validated AUC in the inner 20-fold cross-validation. For ridge regression and the regularization parameter λ, we quantified this variability using standard deviation, with smaller standard deviation suggesting greater stability. For random forest and minimum node size, we assessed stability using the frequency at which the fixed minimum node size was selected in the inner cross-validation. If one grid value was selected at high frequency, we considered there to be stability in minimum node size selected.

Results of the fixed hyperparameter investigation (Section 3.6.1, Section 3.6.2) indicated that tuning parameters could be fixed for all combinations of scenarios of interest and sample size. The fixed tuning parameters selected are provided in Tables 3.4 and 3.5.

Ridge regression

The distribution of regularization parameters, λ, selected by inner 20-fold cross-validation displayed a strong right skew when the training set size was small but became more normally distributed as sample size increased. Additionally, the regularization parameters selected became more stable (smaller standard deviation) as sample size increased and were more stable for larger event rates and a stratified sampling strategy (Figure 3.3, note the $\log_{10}$ scaled x-axis). The standard deviation of the cross-validated AUC also decreased as sample size increased and was smaller for larger event rates (Figure 3.4).

For the larger training set sizes ($n_{\text{train}} = 1 \times 10^5$ and $n_{\text{train}} = 1 \times 10^2$), with 10 replications of the nested cross-validation procedure (Section 3.6.1) the standard deviations of the cross-validated AUCs were small, with SD(AUC) < 0.01 for all but the naive sampling strategy and smallest event rate ($R_0/2$) for $n_{\text{train}} = 1 \times 10^5$. For this combination of rate, sampling

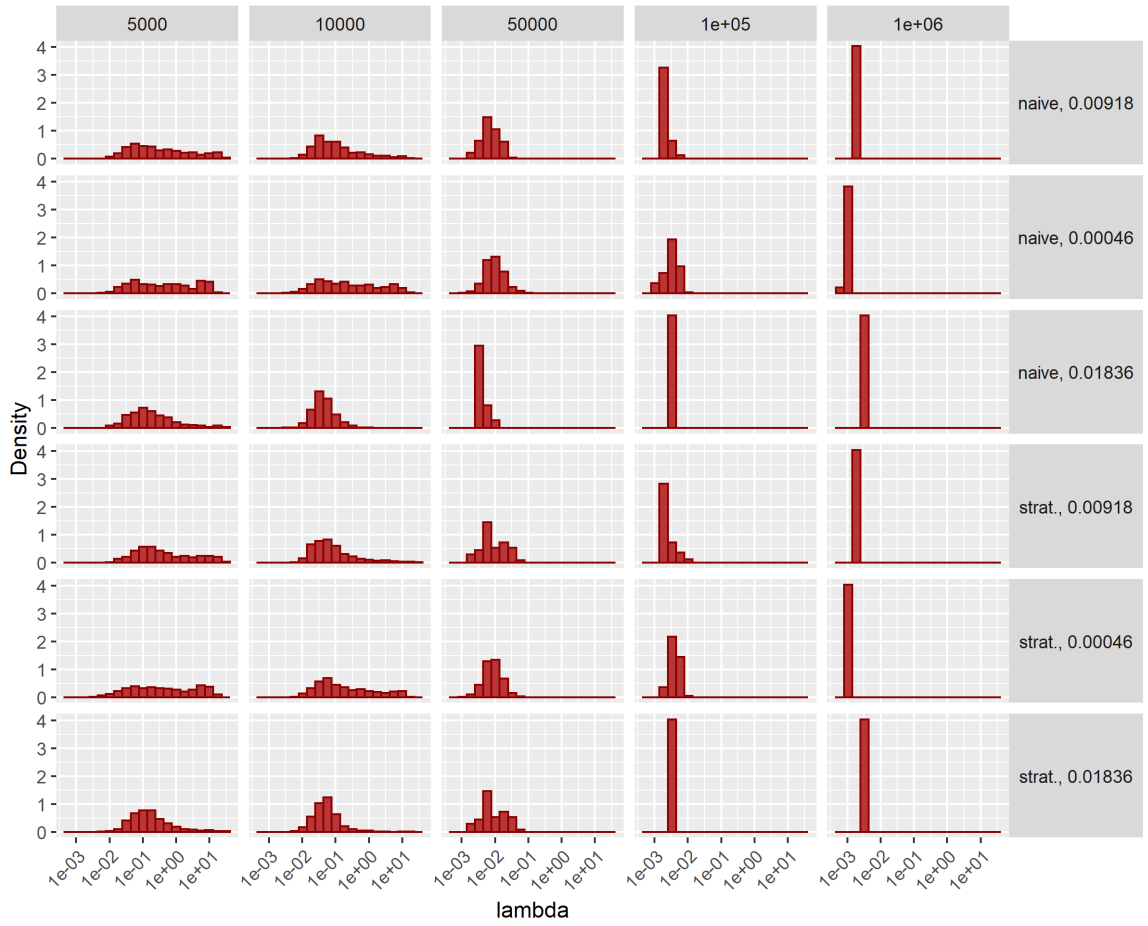

Figure 3.3: Regularization parameters λ selected in the inner folds of the nested cross-validation procedure replicates.

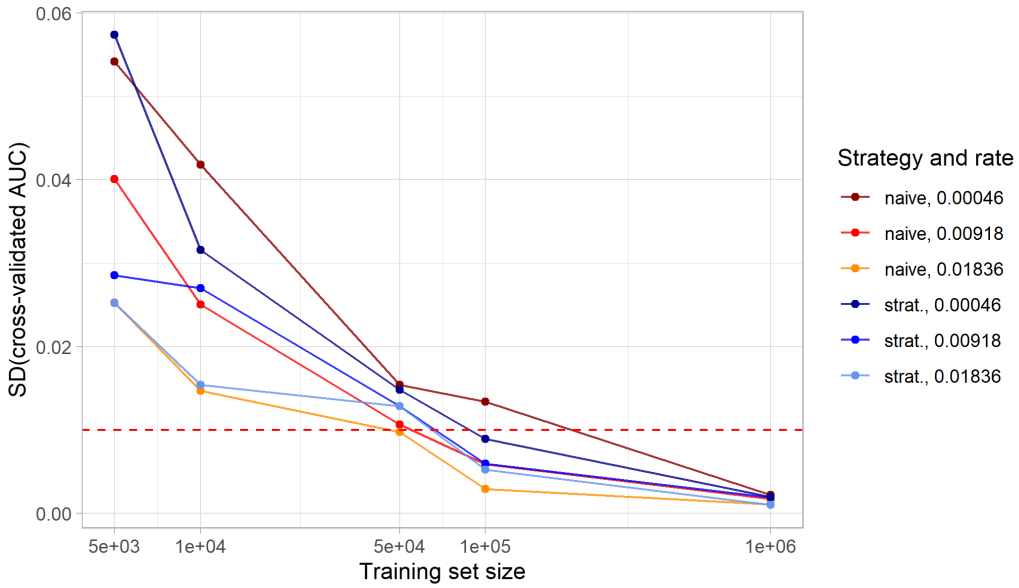

Figure 3.4: Standard deviation of 10-fold cross-validated AUC across nested cross-validation procedure replicates, ridge regression.

strategy, and training set size the standard deviation of AUC after ten replications was 0.013. However, for all combinations of rate, sampling strategy, and these two largest training set sizes the regularization parameters selected were stable. With a training set size of 1×10^5, we found $\text{SD}(\lambda) < 0.002$ for all sampling strategies and rate combinations. With a training set size of 1×10^6, we observed $\text{SD}(\lambda) < 5 \times 1^{-5}$ for all sampling strategy and rate combinations. Thus, for these training set sizes, the fixed tuning parameter was selected by taking the median of 100 selected regularization parameters from ten replications of the nested cross-validation procedure.

For a training set size of 5×10^4, with 10 replications of the pilot simulation (Section 3.6.2) the standard deviation of the cross-validated AUC was less than 0.01, or the standard deviation of the selected regularization parameters was less than 0.01, for the two larger event rates (R_0 and $2R_0$). For the smallest event rate, $R_0/2$, and both sampling strategies after 100 replications of nested cross-validation the standard deviation of the selected regularization parameters was small (near or below 0.01). Thus, for this training set size the fixed regularization parameters were selected by taking the median of 100 selected

regularization parameters from ten iterations of the nested cross-validation procedure for all but the smallest event rate. For the smallest event rate, the fixed regularization parameters were selected by taking the median of 1000 selected regularization parameters from 100 replications of the nested cross-validation procedure.

For the smallest training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$), after 100 replications of the pilot simulation the standard deviation of the cross-validated AUC and the standard deviation of the regularization parameters selected remained above 0.01 for all rates and sampling strategies. However, the results of the pilot simulation indicated that using the fixed tuning parameter did not result in a meaningfully lower cross-validated AUC compared to, in each separate replication, using 20-fold cross-validation on the sampled training set to select the hyperparameter used. The mean difference in cross-validated AUC, $\text{AUC}_{\text{cv}} - \text{AUC}_{\text{fixed}}$, was -0.0006, 0.0014, and -0.0022 for *Pilot Sim. #3, #2* and *#1* (Table 3.3), respectively.

For all training set sizes, we concluded that fixed hyperparameters could be used. For the smallest training set sizes, fixed regularization parameters λ were selected by taking the median of the 1000 selected regularization parameters from 100 replications of nested cross-validation. See Table 3.4 for the values of the chosen hyperparameters.

Random Forest

Stability of the minimum node size selected increased as the training set size increased. The standard deviation of the cross-validated AUC was smaller for larger event rates and, for smaller training set sizes, was smaller with a stratified sampling strategy compared to a naive sampling strategy (Figure 3.5). In all cases, it was possible to select a minimum node size by taking the mode of 100 minimum node sizes selected across 10 replications of the nested cross-validation procedure.

For the larger training set sizes ($n_{\text{train}} = 5 \times 10^4$, 1×10^5, and 1×10^6) the standard deviation of the cross-validated AUC was small (near or below 0.01) and the minimum node size grid values selected were very stable. For these larger training set sizes, over 10 replications of the nested cross-validation, the same minimum node size was selected in

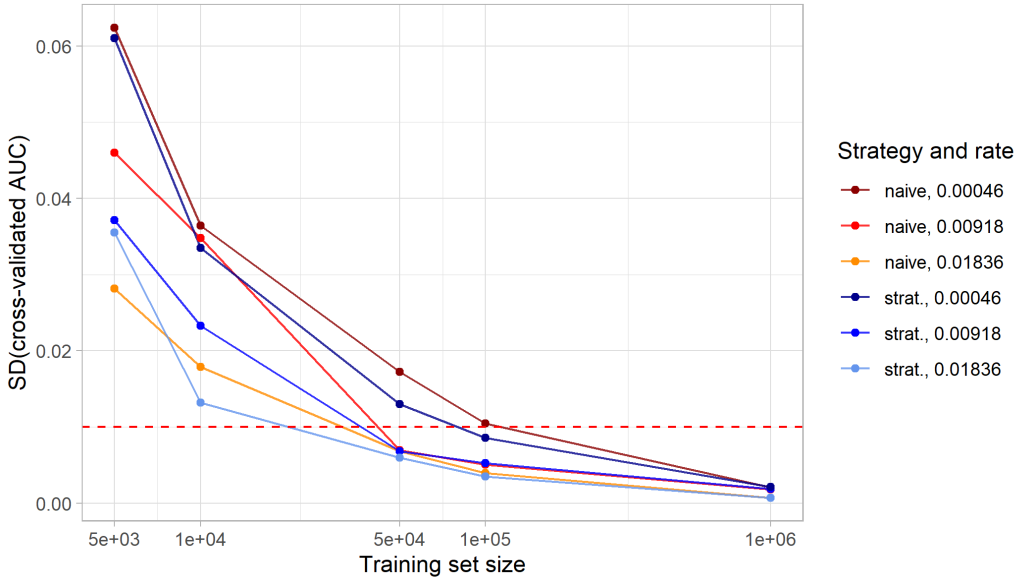

Figure 3.5: Standard deviation of 10-fold cross-validated AUC across nested cross-validation procedure replicates, random forest.

every round of inner cross-validation. For the smaller training set sizes, there was more variability across rounds of inner cross-validation in the minimum node size that maximized cross-validated AUC and the standard deviation of the outer-fold cross-validated AUC was larger. The standard deviation of this outer-fold cross-validated AUC ranged from 0.013 to 0.036 for a training set size of 1×10^4 and from 0.028 to 0.062 for a training set size of 5×10^3. However, using cross-validation to select a minimum node size did not meaningfully increase cross-validated AUC compared to using a fixed minimum node size. The mean difference in cross-validated AUC, $\text{AUC}_{\text{cv}} - \text{AUC}_{\text{fixed}}$, was 0.0016, 0.0023, and 0.0079 for for *Pilot Sim. #3, #2* and both fixed minimum sizes of *Pilot Sim. #1* (Table 3.3), respectively.

3.7 Details of the simulation procedure

For each replication of the main simulation procedure, a training set of size n_{train} and a test set of size 1 million were randomly sampled from the home data set (Section 3.4). The training set was then divided into ten folds using the specified sampling strategy (naive or stratified sampling), and 10-fold cross-validation performed to estimate AUC, Brier score,

Sampling strategy and rate	Training set size				
	5×10^3	1×10^4	5×10^4	1×10^5	1×10^6
naive, $R = R_0/2$	0.398191	0.176459	0.008997	0.003368	0.000866
naive, $R = R_0$	0.210478	0.074001	0.007241	0.001757	0.001645
naive, $R = R_0 \times 2$	0.139739	0.039777	0.003381	0.002935	0.002968
stratified, $R = R_0/2$	0.368710	0.112562	0.007737	0.004101	0.000887
stratified, $R = R_0$	0.271674	0.056800	0.007044	0.001855	0.001650
stratified, $R = R_0 \times 2$	0.137977	0.049776	0.003318	0.002995	0.002987

Table 3.4: Selected values of fixed hyperparameter λ for ridge regression.

Sampling strategy and rate	Training set size				
	5×10^3	1×10^4	5×10^4	1×10^5	1×10^6
naive, $R = R_0/2$	100	1000	1000	1000	10000
naive, $R = R_0$	100	100	1000	1000	10000
naive, $R = R_0 \times 2$	100	100	1000	1000	10000
stratified, $R = R_0/2$	100	1000	1000	1000	10000
stratified, $R = R_0$	100	100	1000	1000	10000
stratified, $R = R_0 \times 2$	100	100	1000	1000	10000

Table 3.5: Selected fixed minimum node sizes for random forest.

accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), F_1 score, and $F_0.5$ score. The metrics requiring predicted events, not predicted probabilities, (all except AUC and Brier score) were calculated using thresholds of 90^{th}, 95^{th}, and 99^{th} percentiles of predicted probabilities from the training folds. For example, for the 90^{th} percentile threshold and hold-out fold k, a cutoff value corresponding to the 90^{th} percentile of predicted probabilities, $p_{k,90\text{th}}$ was determined using the predicted probabilities from the observations in the training folds. Then, predicting on the hold-out fold, all observations with predicted probabilities of greater than $p_{k,90\text{th}}$ were predicted events, and all observations with predicted probabilities of less than or equal to $p_{k,90\text{th}}$ were predicted non-events. We saved the hold-out fold predicted probabilities for bootstrapping later in the simulation procedure. We also obtained influence curve based confidence intervals for cross-validated AUC using the `cvAUC` package [31].

Under a naive sampling strategy, and especially at the smaller training set sizes and

Training set size	Number of trees
5×10^3	250
1×10^4	250
5×10^4	250
1×10^5	100
1×10^6	50

Table 3.6: Number of trees selected for each training set size, random forest.

event rates, on occasion folds had no events. These folds without events were then excluded during cross-validation, since at least one true event was required to obtain an estimate of the hold-out fold AUC, and AUC was the primary prediction performance metric of interest.

We next fit the algorithm on the entire training set to obtain a fitted model and corresponding thresholds of predicted probabilities $p_{90\text{th}}$, $p_{95\text{th}}$, and $p_{99\text{th}}$. These thresholds of predicted probabilities for the model fit on the entire training set and the hold-out fold predicted probabilities from the cross-validation procedure were used to obtain 95% bootstrap confidence intervals for all prediction performance metrics. Two distinct bootstrapping procedures were utilized: one in which re-sampling and metric calculation was performed *within* each cross-validation fold separately, and then the metrics averaged over the folds; and another in which the hold-out fold predicted probabilities were first aggregated and the re-sampling and metric calculation performed *across* the folds. We will henceforth refer to the former bootstrapping procedure as the *within-fold bootstrap* and the latter as the *across-fold bootstrap*. For both procedures, 500 bootstrap replications were used. Using these 500 replications, for each of the two procedures and each of the prediction metrics, both a Wald-type 95% CI and a percentile-type 95% CI were calculated. Thus, for each prediction metric four bootstrap 95% confidence intervals were obtained: across-fold Wald, across-fold percentile, within-fold Wald, and within-fold percentile. Further details of the bootstrapping procedure can be found in subsection 3.7.1.

Finally, we used the model fit on the entire training set and the corresponding thresholds of predicted probabilities $p_{90\text{th}}$, $p_{95\text{th}}$, and $p_{99\text{th}}$ to predict on the test set. We used these test set predictions and the true outcomes to calculate test set AUC, Brier score, accuracy,

specificity, sensitivity, PPV, NPV, F_1 score, and F_2 score calculated at the 90^{th}, 95^{th}, and 99^{th} percentile thresholds.

Overall, for each replication of the main simulation procedure we obtained and saved the following: 10-fold cross-validated prediction metrics and CIs for CV-AUC (obtained on the training set); across-fold 95% Wald-type bootstrap CIs, across-fold 95% percentile-type bootstrap CIs, within-fold 95% Wald-type bootstrap CIs, within-fold 95% percentile-type bootstrap CIs (all obtained on the training set); and test-set prediction metrics. Additionally, for the two smallest training set sizes $n_{\text{train}} = 5 \times 10^3$ and $n_{\text{train}} = 1 \times 10^4$ we performed ten-by-ten cross-validation and did bootstrapping based on this ten-by-ten cross-validation. Thus, for these two smallest training set sizes we obtained cross-validated prediction metrics and bootstrap CIs for both one round of cross-validation and for ten-by-ten cross-validation. Further details on this ten-by-ten cross-validation and bootstrapping can be found in subsection 3.7.2.

3.7.1 Bootstrapping

Bootstrapping was done with the hold-out fold predicted probabilities from the cross-validation procedure and, for those metrics calculated using predicted events, the 90^{th}, 95^{th}, and 99^{th} percentile thresholds of predicted probabilities from the model fit on the entire training set. When there were folds without events, as occurred occasionally under a naive sampling strategy, we used only the hold-out fold predicted probabilities from folds with events.

For the within-fold bootstrapping procedure, for each bootstrap replication, we first re-sampled with replacement within each cross-validation fold separately to obtain a bootstrap sample for each fold. With each fold's bootstrap sample, we then calculated the fold-specific prediction metrics, using the percentile thresholds to obtain predicted events from the predicted probabilities as needed. We then averaged over the folds, in a manner analogous to cross-validation, to obtain the bootstrap prediction metrics. For the across-fold bootstrapping procedure, for each bootstrap replication, we first aggregated the hold-out fold predicted probabilities. We then re-sampled the predicted probabilities with replacement to

obtain the bootstrap sample. From this bootstrap sample, we then calculated the bootstrap prediction metrics, using the percentile thresholds as needed.

Five-hundred replications of the above procedure resulted in two sets of bootstrap values, each of size 500, for each prediction metric m:

$$M_{\text{within}} = \{m_{1,\text{within}}, m_{2,\text{within}}, \ldots, m_{500,\text{within}}\}, \text{ and}$$
$$M_{\text{across}} = \{m_{1,\text{across}}, m_{2,\text{across}}, \ldots, m_{500,\text{across}}\}.$$

The across- and within-fold Wald-type bootstrap 95% confidence intervals were then calculated, respectively, as

$$(\bar{M}_{\text{within}} - 1.96 \cdot \text{SD}(M_{\text{within}}), \bar{M}_{\text{within}} + 1.96 \cdot \text{SD}(M_{\text{within}})), \text{ and}$$
$$(\bar{M}_{\text{across}} - 1.96 \cdot \text{SD}(M_{\text{across}}), \bar{M}_{\text{across}} + 1.96 \cdot \text{SD}(M_{\text{across}})),$$

where $\bar{M}$ represents the mean of the 500 bootstrap values. The across- and within-fold percentile bootstrap 95% CIs were calculated as

$$(Q_{2.5\%}(M_{\text{within}}), Q_{97.5\%}(M_{\text{within}})), \text{ and}$$
$$(Q_{2.5\%}(M_{\text{across}}), Q_{97.5\%}(M_{\text{across}})),$$

where $Q_x(M)$ is the x quantile of the 500 bootstrap values.

3.7.2 Ten-by-ten cross-validation and bootstrap confidence intervals

For the two smallest training set sizes, 5×10^3 and 1×10^4, we conducted ten-by-ten cross validation to estimate the prediction metrics. In ten-by-ten cross-validation, ten rounds of ten-fold cross-validation are performed, with a different random split of the training data into ten folds each time, resulting in ten (one for each round) cross-validated estimates of each metric. Then, averaging over the cross-validation rounds yields a single average cross-validated estimate of the metric.

In performing ten-by-ten cross-validation we obtained ten sets of hold-out probabili-

ties, one for each cross-validation round. Using these ten sets of holdout probabilities, we obtained *ten-by-ten bootstrap confidence intervals*. For each cross-validation round, we performed the 500 replications of the bootstrap procedure, resulting in 5,000 bootstrap estimates of each metric, from which we calculated Wald-type and percentile bootstrap confidence intervals.

3.8 Investigating the variability introduced by sampling a test set

Although the size of the test set is large (1 million), sampling the test set from the full data set introduces sampling variability into the iteration-specific test set prediction metric value. To ascertain the size of this variability—distinct from the variability introduced by sampling the training set—for each algorithm we calculated the standard deviation, IQR, and range of 1,000 test set AUC values from using the same model to predict on independently sampled test sets. We used an event rate of $R_0 = 0.0092$, and trained the models on a training set of size 1 million. For the ridge logistic regression and random forest algorithms, we used the fixed hyperparameters corresponding to this rate, training set size, and a naive sampling strategy.

3.9 Monte Carlo standard error

The Monte Carlo error (MCE), a measure of between-simulation variability, is formally defined as the standard deviation of the Monte Carlo estimator $\widehat{\varphi}_r$, where r is the number of replicates:

$$\text{MCE}(\widehat{\varphi}_r) = \sqrt{\text{Var}(\widehat{\varphi}_r)}. \quad [29]$$

We estimated the MCE of the empirical bias and the empirical coverage of 95% CIs using an asymptotic approach. In a simulation with r replicates, for each replicate i with sampled data X_i we obtain the target quantity $\varphi(X_i)$. For empirical coverage, this is an indicator of if the 95% confidence interval covers the test set value. For empirical bias, this is the distance between the test set value and the training set value. The Monte Carlo estimator is then

$$\widehat{\varphi}_r = \frac{1}{r}\sum_{i=1}^{r}\varphi(X_i).$$

Appealing to the strong law of large numbers and the Central Limit Theorem, an estimate of the MCE is

$$\widehat{\mathrm{MCE}}(\widehat{\varphi}_r) = \frac{1}{r}\sqrt{\sum_{i=1}^{r}(\varphi(X_i) - \widehat{\varphi}_r)^2}. \quad [29]$$

Chapter 4

RESULTS: AUC

4.1 Empirical Coverage

4.1.1 Number of events in the training set drives coverage

For random forest and ridge logistic regression, the empirical coverage of the 95% confidence intervals for the iteration-specific test set AUC initially increased as the number of training set events increased, but then decreased for the largest number of events. Figure 4.1 shows this pattern for the percentile-type bootstrap confidence intervals calculated by bootstrap sampling across the aggregated holdout fold predicted probabilities. For both random forest and ridge regression, and for all four types of bootstrap confidence intervals, empirical coverage was between approximately 90% and 95% for all but the training set size and event rate combinations that resulted in the three smallest number of events in the training set, and for a training set size of 1 million. Empirical coverage was worst at a training set size of 1 million for the bootstrap confidence intervals, with coverage of approximately 80-85%. Empirical coverage of the influence-curve based confidence intervals for the cross-validated AUC followed a pattern similar to that of the bootstrap confidence intervals. However, for these influence curve based confidence intervals the lowest empirical coverage occurred at the two smallest training set sizes and was below 80% (Tables 4.1, 4.2).

Empirical coverage of the 95% confidence intervals for test set AUC was generally lower for logistic regression than for random forest or ridge logistic regression. Only for an absolute number of events $n_{\text{event}} = 919$, and $n_{\text{event}} = 1837$—with the exception of the influence curve based confidence intervals at the smallest total number of events, $n_{\text{event}} = 23$—did we observe an empirical coverage of greater than 90%. Additionally, with logistic regression the empirical coverage decreased from the smallest number of events to $n_{\text{event}} = 92$, then increased from $n_{\text{event}} = 184$ to $n_{\text{event}} = 1837$, until decreasing again for the largest number of events ($n_{\text{event}} = 4591$ and $n_{\text{event}} = 9181$) (Figure 4.1, Table, 4.3).

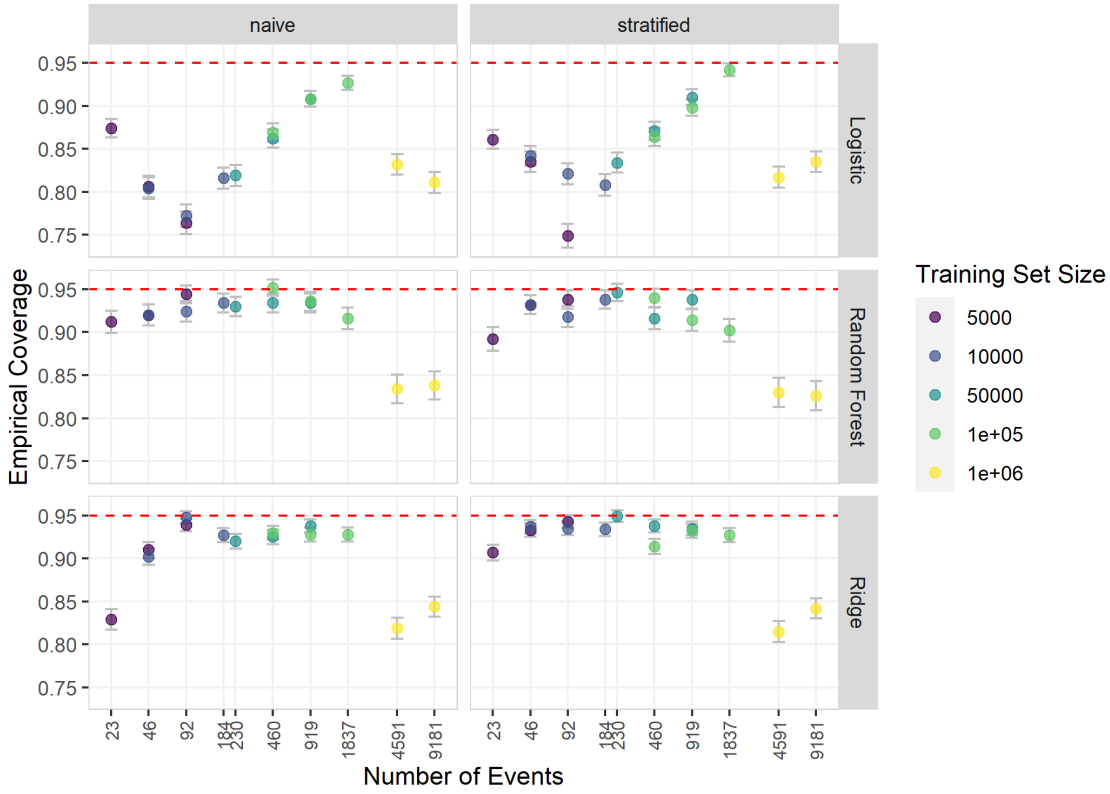

Figure 4.1: Empirical coverage of test set AUC using percentile-type bootstrap confidence intervals versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size, while the red dotted lines represent the nominal 95% coverage level. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

Empirical coverage of the 95% confidence intervals for test set AUC was driven by number of events in the training set, not event rate. In simulations with different training set sizes and event rates, but the same absolute number of events—e.g., $n_{\text{train}} = 50,000$ and $R = 0.0092$ versus $n_{\text{train}} = 100,000$ and $R = 0.0046$, both of which have $n_{\text{event}} = 460$—there was no systematic difference in coverage. That is, in such pairs the combination with a lower event rate did not systematically have worse coverage. And, in most of these pairs, there was no difference in coverage at the level of precision appropriate given the Monte Carlo error present.

Except for the simulations with the smallest number of events in the training set ($n_{\text{event}} = 23, 36,$ and 92), there was no difference—after accounting for Monte Carlo Error—in coverage obtained by using a stratified sampling strategy compared to a naive sampling strategy. When the training set had few events, the empirical coverage depended on the combination of sampling strategy, algorithm, and the the number of events. For example, at $n_{\text{event}} = 23$ a naive sampling strategy resulted in higher empirical coverage in logistic regression and random forest, but lower coverage in ridge regression. At $n_{\text{event}} = 46$, a stratified sampling strategy resulted in higher empirical coverage for all algorithms.

4.1.2 Coverage is similar for the different types of bootstrap confidence intervals

Overall, the empirical coverage for test set AUC for different types of bootstrap confidence intervals (Section 3.7.1) was similar. We did not observe that one procedure for bootstrapping (sampling with replacement across the aggregated holdout fold probabilities versus sampling with replacement within each fold separately) or one type of confidence interval (percentile or Wald) resulted in meaningfully and consistently (across algorithms, training set sizes, sampling strategy, and rates) better coverage than another. We did, however, observe that the influence curve based confidence intervals had worse performance than the bootstrap confidence intervals when the number of events in the training set was small. It was not until $n_{\text{event}} \approx 184$ that the coverage of these influence curve based confidence intervals was similar to that of the bootstrap confidence intervals. The exception to this pattern is the high (99%) empirical coverage observed with the influence curve confidence

			Coverage				
Training set size	Rate	Sampling strategy	Influence	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	0.752	0.904	0.868	0.912	0.87
5,000	0.0046	stratified	0.77	0.88	0.848	0.892	0.854
5,000	0.0092	naive	0.816	0.918	0.884	0.92	0.894
5,000	0.0092	stratified	0.89	0.946	0.938	0.932	0.938
5,000	0.0184	naive	0.902	0.95	0.92	0.944	0.926
5,000	0.0184	stratified	0.924	0.94	0.932	0.938	0.942
10,000	0.0046	naive	0.816	0.916	0.892	0.92	0.9
10,000	0.0046	stratified	0.892	0.934	0.926	0.932	0.936
10,000	0.0092	naive	0.878	0.926	0.92	0.924	0.916
10,000	0.0092	stratified	0.902	0.924	0.922	0.918	0.926
10,000	0.0184	naive	0.928	0.942	0.946	0.934	0.934
10,000	0.0184	stratified	0.934	0.944	0.94	0.938	0.942
50,000	0.0046	naive	0.928	0.93	0.93	0.93	0.93
50,000	0.0046	stratified	0.934	0.948	0.938	0.946	0.936
50,000	0.0092	naive	0.928	0.936	0.932	0.934	0.928
50,000	0.0092	stratified	0.922	0.922	0.922	0.916	0.922
50,000	0.0184	naive	0.93	0.938	0.928	0.934	0.932
50,000	0.0184	stratified	0.942	0.942	0.94	0.938	0.94
100,000	0.0046	naive	0.94	0.95	0.948	0.952	0.954
100,000	0.0046	stratified	0.934	0.934	0.934	0.94	0.934
100,000	0.0092	naive	0.94	0.944	0.936	0.936	0.936
100,000	0.0092	stratified	0.912	0.912	0.91	0.914	0.912
100,000	0.0184	naive	0.922	0.916	0.914	0.916	0.916
100,000	0.0184	stratified	0.906	0.908	0.912	0.902	0.902
1,000,000	0.0046	naive	0.836	0.842	0.83	0.834	0.832
1,000,000	0.0046	stratified	0.84	0.834	0.84	0.83	0.838
1,000,000	0.0092	naive	0.846	0.84	0.85	0.838	0.846
1,000,000	0.0092	stratified	0.83	0.83	0.834	0.826	0.832

Table 4.1: Empirical coverage of 95% confidence intervals for AUC with the random forest algorithm. Colors correspond to coverage.

intervals with logistic regression at the smallest number of training set events (Tables 4.1, 4.2, 4.3). However, the logistic regression influence function based confidence intervals at this smallest number of events were wide (average width, 0.39 and 0.38 for a naive and stratified sampling strategy, respectively) and there was high positive bias (Figure 4.4).

4.1.3 Width of confidence intervals decreases as n_{event} increases

The average width ($\text{CI}_{\text{upper bound}} - \text{CI}_{\text{lower bound}}$) of the 95% confidence intervals decreased as the number of events increased (Figure 4.2). As with coverage, this decrease in width appeared to be driven by the number of events. The average width of the confidence intervals

			Coverage				
Training set size	Rate	Sampling strategy	Influence	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	0.783	0.848	0.884	0.829	0.893
5,000	0.0046	stratified	0.798	0.912	0.878	0.907	0.876
5,000	0.0092	naive	0.837	0.912	0.905	0.91	0.918
5,000	0.0092	stratified	0.898	0.937	0.935	0.933	0.939
5,000	0.0184	naive	0.902	0.938	0.933	0.939	0.936
5,000	0.0184	stratified	0.922	0.943	0.944	0.943	0.942
10,000	0.0046	naive	0.82	0.905	0.889	0.902	0.89
10,000	0.0046	stratified	0.891	0.934	0.926	0.937	0.924
10,000	0.0092	naive	0.925	0.955	0.951	0.948	0.953
10,000	0.0092	stratified	0.906	0.937	0.94	0.935	0.935
10,000	0.0184	naive	0.92	0.933	0.937	0.927	0.932
10,000	0.0184	stratified	0.935	0.942	0.943	0.934	0.944
50,000	0.0046	naive	0.909	0.928	0.927	0.92	0.925
50,000	0.0046	stratified	0.939	0.952	0.947	0.949	0.945
50,000	0.0092	naive	0.925	0.93	0.929	0.925	0.925
50,000	0.0092	stratified	0.94	0.945	0.947	0.938	0.938
50,000	0.0184	naive	0.943	0.944	0.944	0.938	0.942
50,000	0.0184	stratified	0.943	0.939	0.943	0.935	0.943
100,000	0.0046	naive	0.934	0.936	0.937	0.93	0.94
100,000	0.0046	stratified	0.909	0.916	0.914	0.914	0.909
100,000	0.0092	naive	0.934	0.93	0.935	0.928	0.935
100,000	0.0092	stratified	0.93	0.933	0.931	0.932	0.929
100,000	0.0184	naive	0.932	0.934	0.933	0.928	0.931
100,000	0.0184	stratified	0.931	0.928	0.933	0.927	0.924
1,000,000	0.0046	naive	0.824	0.82	0.819	0.819	0.817
1,000,000	0.0046	stratified	0.822	0.82	0.818	0.815	0.812
1,000,000	0.0092	naive	0.857	0.852	0.852	0.844	0.844
1,000,000	0.0092	stratified	0.853	0.851	0.854	0.842	0.851

Table 4.2: Empirical coverage of 95% confidence intervals for AUC with ridge regression. Colors correspond to coverage.

			Coverage				
Training set size	Rate	Sampling strategy	Influence	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	**0.991**	**0.874**	**0.844**	**0.874**	**0.843**
5,000	0.0046	stratified	**0.991**	**0.865**	**0.821**	**0.861**	**0.821**
5,000	0.0092	naive	**0.833**	**0.805**	**0.865**	**0.806**	**0.863**
5,000	0.0092	stratified	**0.828**	**0.828**	**0.831**	**0.835**	**0.832**
5,000	0.0184	naive	**0.777**	**0.776**	**0.816**	**0.764**	**0.818**
5,000	0.0184	stratified	**0.763**	**0.763**	**0.784**	**0.749**	**0.781**
10,000	0.0046	naive	**0.734**	**0.805**	**0.814**	**0.804**	**0.809**
10,000	0.0046	stratified	**0.804**	**0.836**	**0.849**	**0.842**	**0.848**
10,000	0.0092	naive	**0.779**	**0.784**	**0.826**	**0.772**	**0.82**
10,000	0.0092	stratified	**0.816**	**0.825**	**0.84**	**0.821**	**0.833**
10,000	0.0184	naive	**0.819**	**0.818**	**0.836**	**0.816**	**0.835**
10,000	0.0184	stratified	**0.813**	**0.82**	**0.824**	**0.808**	**0.808**
50,000	0.0046	naive	**0.823**	**0.829**	**0.844**	**0.819**	**0.832**
50,000	0.0046	stratified	**0.834**	**0.845**	**0.847**	**0.834**	**0.835**
50,000	0.0092	naive	**0.874**	**0.871**	**0.882**	**0.862**	**0.871**
50,000	0.0092	stratified	**0.88**	**0.882**	**0.886**	**0.871**	**0.88**
50,000	0.0184	naive	**0.912**	**0.908**	**0.914**	**0.908**	**0.91**
50,000	0.0184	stratified	**0.92**	**0.921**	**0.916**	**0.91**	**0.914**
100,000	0.0046	naive	**0.878**	**0.873**	**0.888**	**0.869**	**0.88**
100,000	0.0046	stratified	**0.874**	**0.874**	**0.878**	**0.864**	**0.872**
100,000	0.0092	naive	**0.914**	**0.909**	**0.912**	**0.908**	**0.909**
100,000	0.0092	stratified	**0.897**	**0.899**	**0.9**	**0.898**	**0.891**
100,000	0.0184	naive	**0.94**	**0.935**	**0.94**	**0.927**	**0.939**
100,000	0.0184	stratified	**0.941**	**0.938**	**0.942**	**0.942**	**0.935**
1,000,000	0.0046	naive	**0.841**	**0.843**	**0.836**	**0.832**	**0.831**
1,000,000	0.0046	stratified	**0.822**	**0.818**	**0.817**	**0.817**	**0.816**
1,000,000	0.0092	naive	**0.819**	**0.816**	**0.819**	**0.811**	**0.811**
1,000,000	0.0092	stratified	**0.841**	**0.839**	**0.842**	**0.835**	**0.837**

Table 4.3: Empirical coverage of 95% confidence intervals for AUC with logistic regression. Colors correspond to coverage.

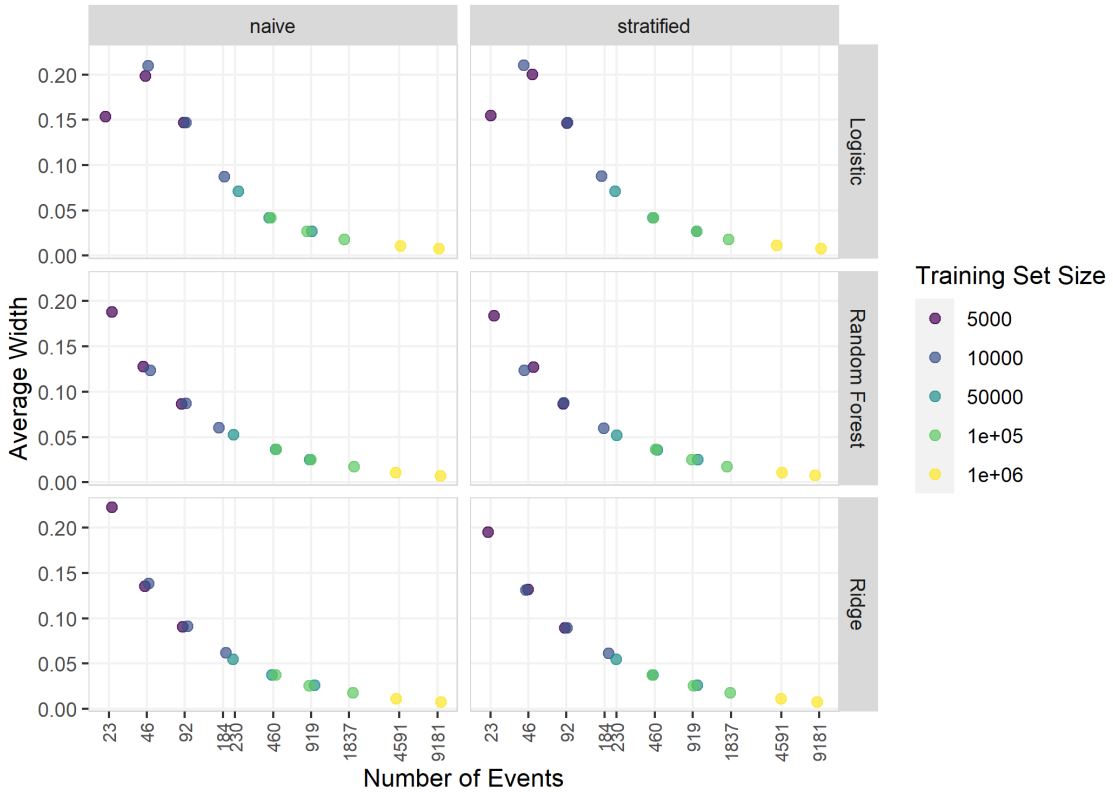

Figure 4.2: Average width of percentile-type bootstrap confidence intervals for AUC versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

was similar between random forest and ridge regression, and for these two algorithms the decrease in width with number of training set events was monotone. When the number of events was smaller (approximately $n_{\text{event}} < 460$), the average width of the confidence intervals was wider with logistic regression than with random forest or ridge regression. At larger number of events the widths were more similar between the three algorithms. Additionally, with logistic regression the average width of the confidence intervals at the smallest number of events, $n_{\text{event}} = 23$, was smaller than the average width of the confidence intervals at the next smallest number of events, $n_{\text{event}} = 46$.

At the largest number of events, corresponding to a training set size of 1 million, the confidence intervals were very narrow for all algorithms, with an average width of approximately 0.01. For a training set size of 100,000, the average width of the confidence intervals was approximately 0.04, 0.03, and 0.02 for event rates $R = 0.0046$ ($n_{\text{event}} = 460$), $R = 0.0092$ ($n_{\text{event}} = 919$), and $R = 0.0184$ ($n_{\text{event}} = 1837$), respectively.

4.1.4 Averaging test-set AUC increases empirical coverage

We also calculated the empirical coverage of the 95% confidence intervals for the average test set AUC, where we averaged over the iteration-specific test set AUCs from the simulation replicates of a given training set size, event rate, sampling strategy, and algorithm combination (either 1,000 or 500). This empirical coverage of the average test set AUC is

$$\frac{1}{r}\sum_{i=1}^{r} I_{95\%\text{CI}}\left(\tfrac{1}{r}\textstyle\sum_{i=1}^{r}\text{AUC}_{i,\text{test}}\right),$$

in contrast to the empirical coverage of the iteration-specific test set AUC

$$\frac{1}{r}\sum_{i=1}^{r} I_{95\%\text{CI}}\left(\text{AUC}_{i,\text{test}}\right),$$

where r is the number of replicates. Note that while the algorithms and, if applicable, hyperparameters remained constant between simulation iterations, the models differed because they were trained on different randomly sampled training sets.

At the smaller training set sizes and number of events, there was little difference in the empirical coverage of the iteration-specific test set AUC and the average test set AUC. However, as the training set size increased the empirical coverage of the average test set AUC and the iteration-specific test set AUC diverged. While the coverage of the average test set AUC remained at or above 95% when $n_{\text{train}} = 1$ million and $n_{\text{train}} = 100,000$, the coverage of the iteration-specific test set did not. At a training set size of 1 million, there was an approximately 10 percentage point difference in coverage. At a training set size of 100,000, the coverage of the average test set AUC was slightly higher than the coverage of the iteration-specific test set AUC. For logistic regression, there was a difference in empirical

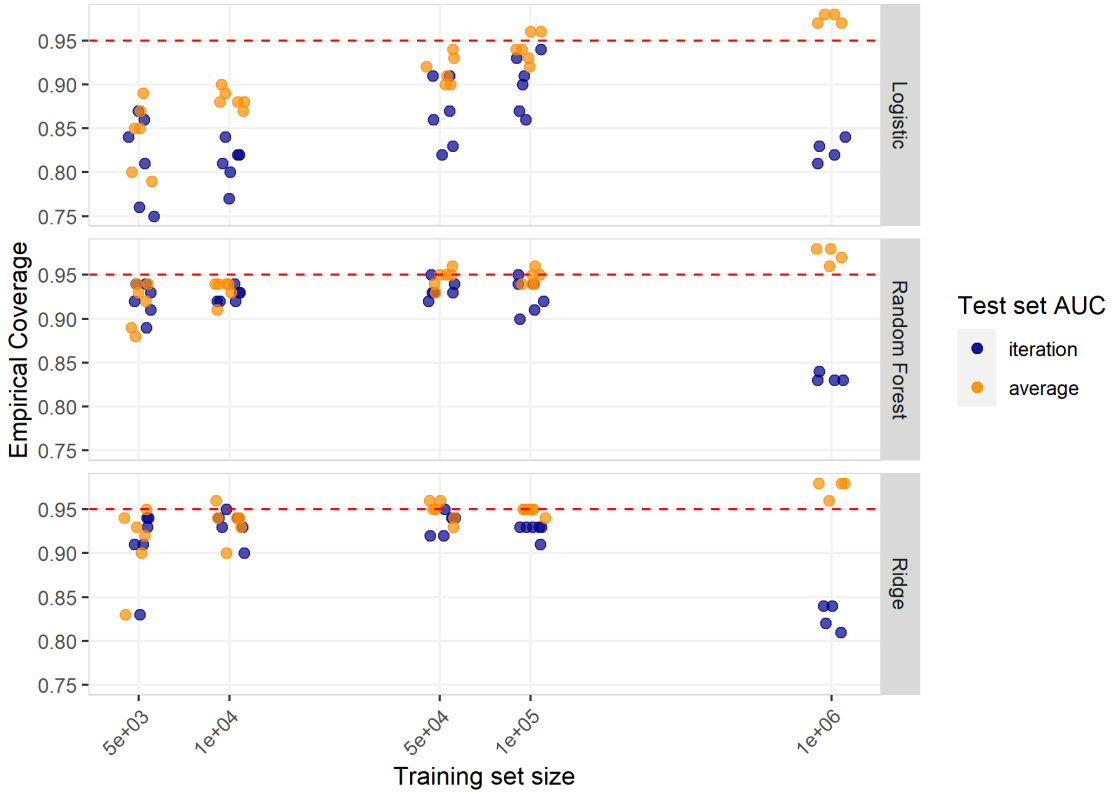

Figure 4.3: Empirical coverage of iteration-specific average (over iterations) test set AUC versus number of events in the training set. Coverage is for bootstrap percentile-type intervals, where bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Both sampling strategies (naive and stratified) are represented. Colors correspond to coverage of the iteration-specific or average test set values, while the red dotted lines represent the nominal 95% coverage level. Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

coverage for all but the smallest training set size of 5,000 (Figure 4.3).

4.2 Empirical Bias

4.2.1 Bias decreases as n_{event} increases

The empirical bias of the cross-validated AUC estimate for test set AUC trended toward zero as the number of events in the training set increased. Similar to the pattern for empirical coverage of the 95% confidence intervals, the empirical bias for AUC appeared to be driven by the number of events in the training set. The empirical bias was almost always negative, indicating that the cross-validated AUC estimate was on average lower than the iteration-specific test set AUC for most combinations of training set size, event rate, sampling strategy, and algorithm. A notable exception is the positive bias observed for logistic regression at $n_{\text{event}} = 23$ and for random forest and ridge regression at $n_{\text{event}} = 23$, $n_{\text{event}} = 46$, and $n_{\text{event}} = 92$ for some combinations of training set size, sampling strategy, and rate. However, for ridge regression and random forest the Monte Carlo standard error for the empirical bias was relatively large relative to the size of the bias.

The size of the empirical bias for test set AUC was similar for the ridge logistic regression and random forest algorithms. There was greater empirical bias with logistic regression. At the smaller training set size, the empirical bias observed with logistic regression was an order of magnitude larger than the empirical bias observed with ridge regression or random forest. When the training set had a large number of events (approximately $n_{\text{event}} > 1,000$), the empirical bias was very close to zero for all algorithms. For example, with a training set size of 1 million we had |empirical bias| ≤ 0.0005 for all algorithms, and with a training set size of 100,000 we had |empirical bias| ≤ 0.001 for random forest and ridge regression and |empirical bias| ≤ 0.005 for logistic regression.

4.2.2 Variability of $AUC_{CV} - AUC_{test}$ decreases as n_{event} increases

As the number of events in the training set increased—so, for a given training set size, as the event rate increased—there was less variability in the value of $\text{AUC}_{\text{CV}} - \text{AUC}_{\text{test}}$ across simulation iterations (Figure 4.5). With a smaller number of events in the training set

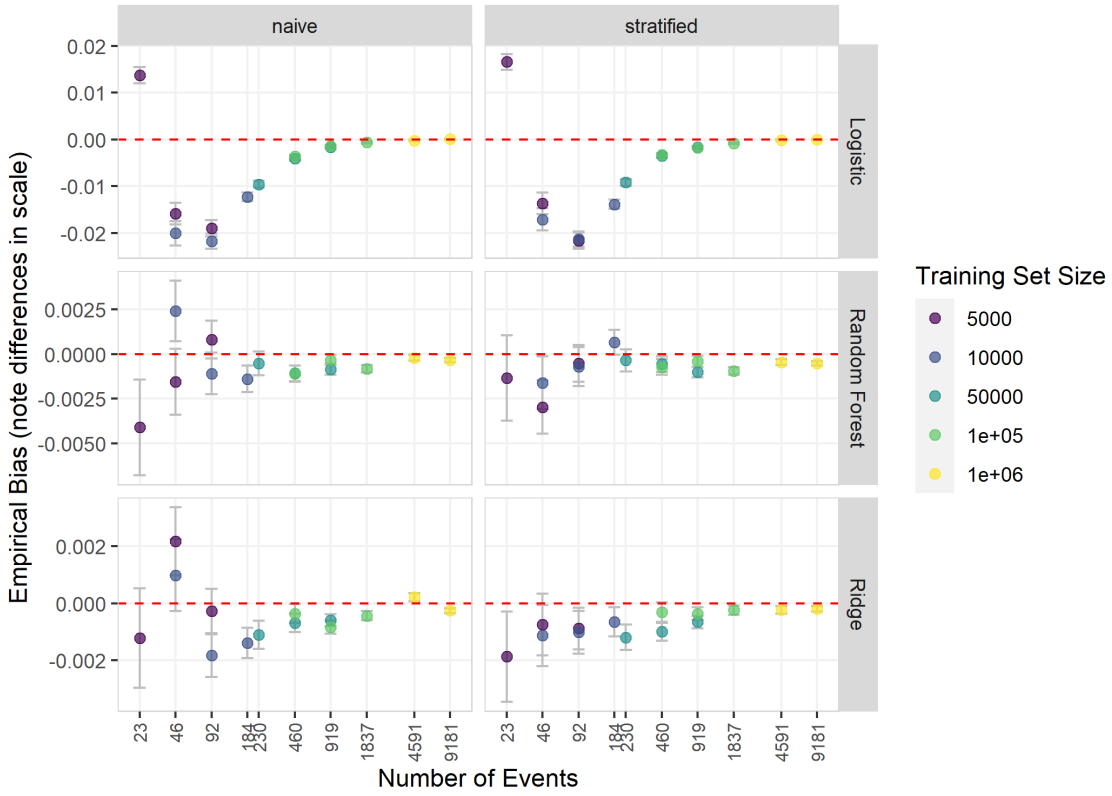

Figure 4.4: Empirical bias for AUC ($\text{AUC}_{\text{CV}} - \text{AUC}_{\text{test}}$) versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

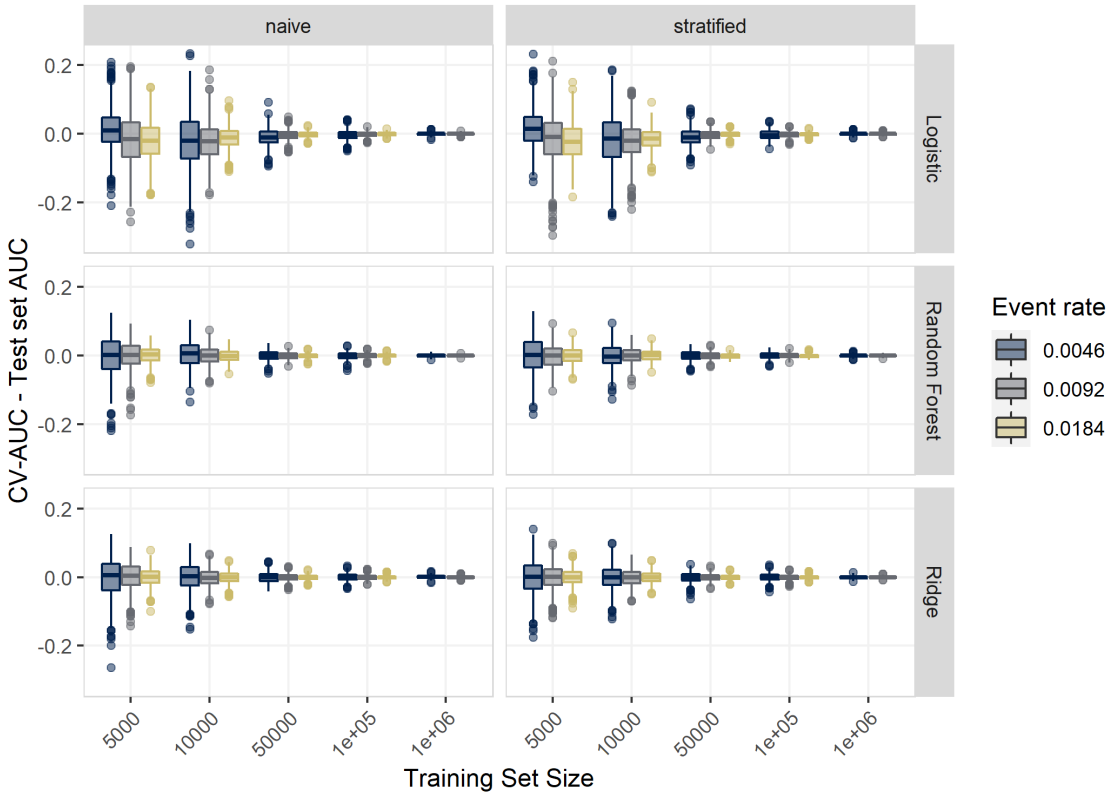

Figure 4.5: Boxplots of iteration-specific bias for AUC ($\text{AUC}_{\text{CV}} - \text{AUC}_{\text{test}}$), stratified by event rate, versus training set size. Colors correspond to event rate. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

($n_{\text{event}} < 200$), the absolute difference between the cross-validated AUC estimate and the test set AUC was, in many simulation iterations, greater than 0.1. While the empirical bias for AUC was mostly negative, for any one simulation replicate the cross-validated AUC could either underestimate or overestimate the test set AUC. The greatest variability in $\text{AUC}_{\text{CV}} - \text{AUC}_{\text{test}}$ occurred with logistic regression at the smaller training set sizes.

4.2.3 Absolute bias scaled by $\sqrt{n_{train}}$ does not trend to zero

The coverage of the 95% confidence intervals is based on the bias scaled (i.e., multiplied) by the square root of the training set size going to zero as the square root of the training set size

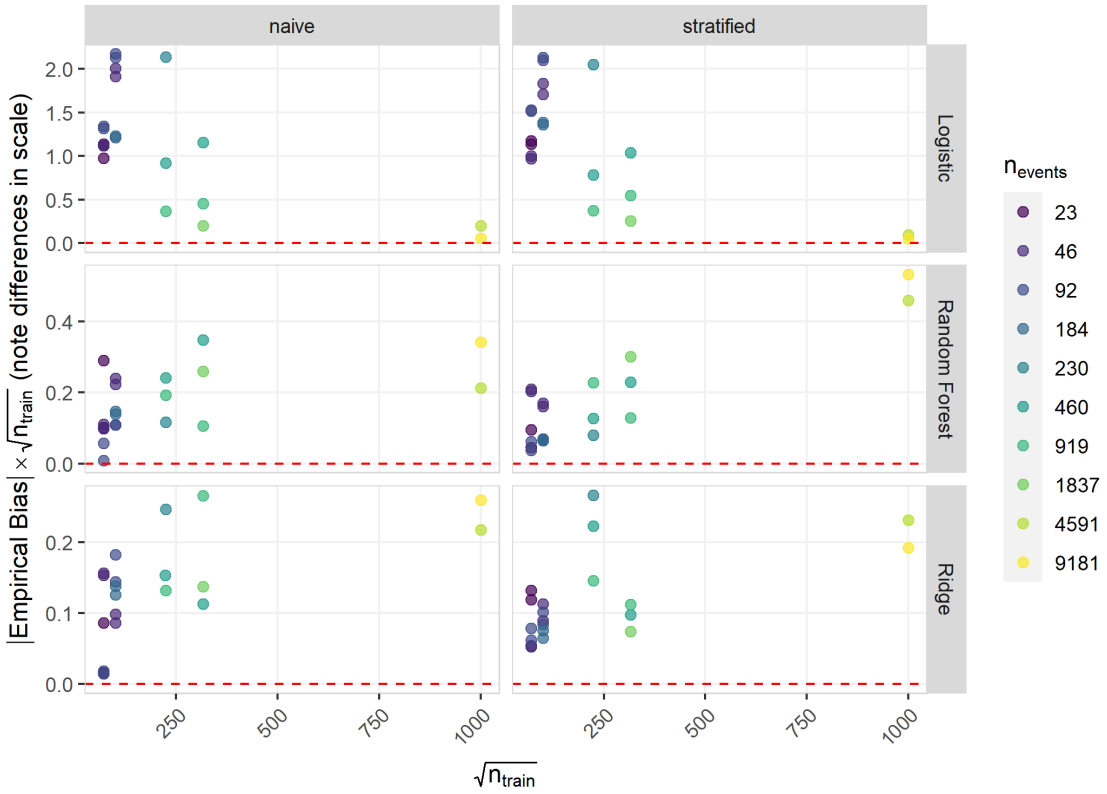

Figure 4.6: Absolute empirical bias scaled by $\sqrt{n_{\text{train}}}$ versus $\sqrt{n_{\text{train}}}$. Colors represent the number of events in the training set size. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

increases. While the empirical bias for AUC trended toward zero as the number of events in the training set increased, for random forest and ridge regression the absolute value of the empirical bias scaled by the square root of the training set size did not trend to zero with the square root of training set size (Figure 4.6). Instead, the scaled absolute empirical bias increased or remained flat. In contrast, for logistic regression the scaled absolute empirical bias did trend to zero with the square root of the training set size. However, at smaller training set sizes the scaled absolute empirical bias was larger for logistic regression than for ridge regression or random forest.

4.3 Empirical Mean Squared Error

4.3.1 Mean squared error decreases as n_{event} increases

The empirical mean squared error (MSE) trended quickly to zero as the number of events in the training set increased (Figure 4.7). Like with bias and confidence interval coverage, the MSE appeared to be driven by number of events, not event rate. For a given number of events in the training set, there was no systematic difference in MSE by training set size or event rate. The decrease in empirical MSE as n_{event} increased was monotone for the ridge regression and random forest algorithms, while for logistic regression the decrease in MSE was montonone for $n_{\text{event}} \geq 46$. The MSE was less than 0.001 when $n_{\text{event}} \geq 184$ for all algorithms and both a naive and stratified sampling strategy. Except for random forest and ridge regression at the two smallest number of training set events, $n_{\text{event}} = 23$ and $n_{\text{event}} = 46$, there was little difference in MSE between a naive and stratified sampling strategy. At these small number of events, the MSE was lower with a stratified sampling strategy.

4.4 Variability and Trends in Cross Validated and Test Set AUC

Variability in the cross-validated AUC estimate across simulation iterations decreased as the number of events in the training set increased (Figure 4.8). For a given training set size, there was less variability at larger event rates (more events in the training set). The mean cross-validated AUC also increased with the number of events in the training set. There was little overall difference in the variability of cross-validated AUC comparing a naive and stratified sampling strategy for creating cross-validation folds.

At the three smallest training set sizes (n_{train} = 5,000, 10,000, and 50,000) there was much greater variability in the cross-validated AUC estimates from logistic regression compared to the random forest and ridge logistic regression algorithms. On average, the cross-validated AUC estimates from logistic regression were lower than those from the random forest or ridge regression algorithms. The variability and values of cross-validated AUC estimates were comparable for random forest and ridge logistic regression.

Similarly, variability in the test set AUC across simulation iterations decreased as the

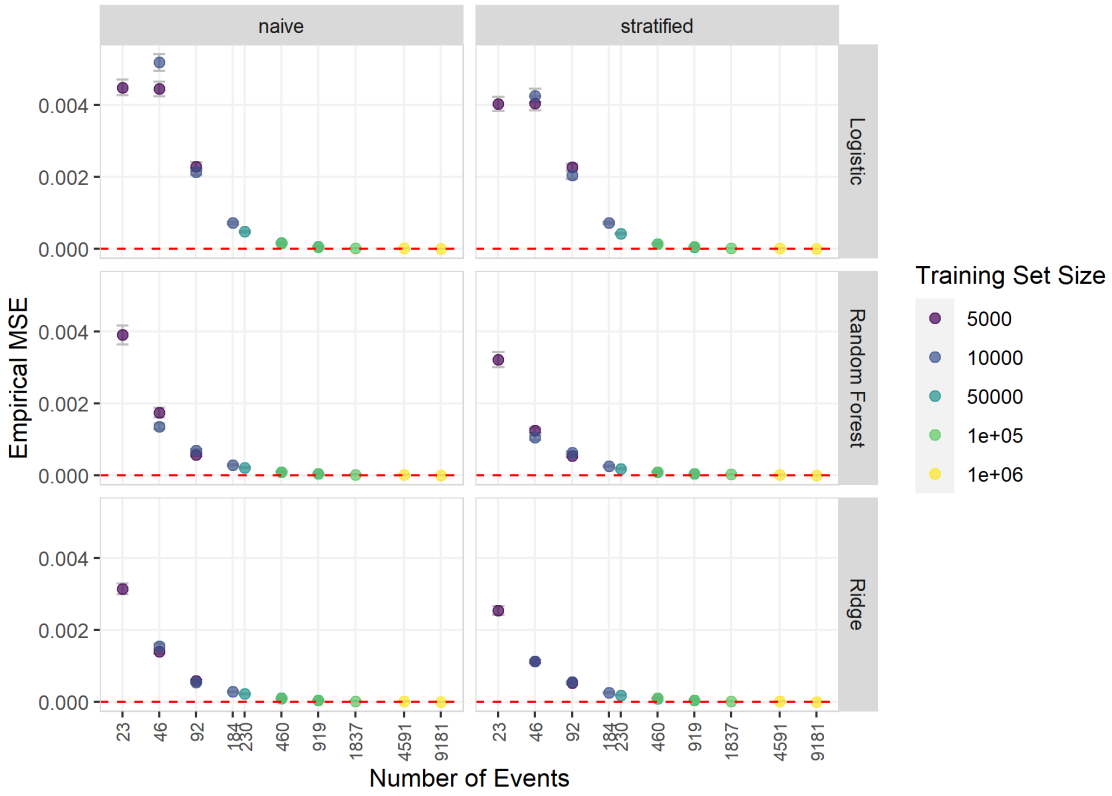

Figure 4.7: Empirical mean squared error for AUC versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

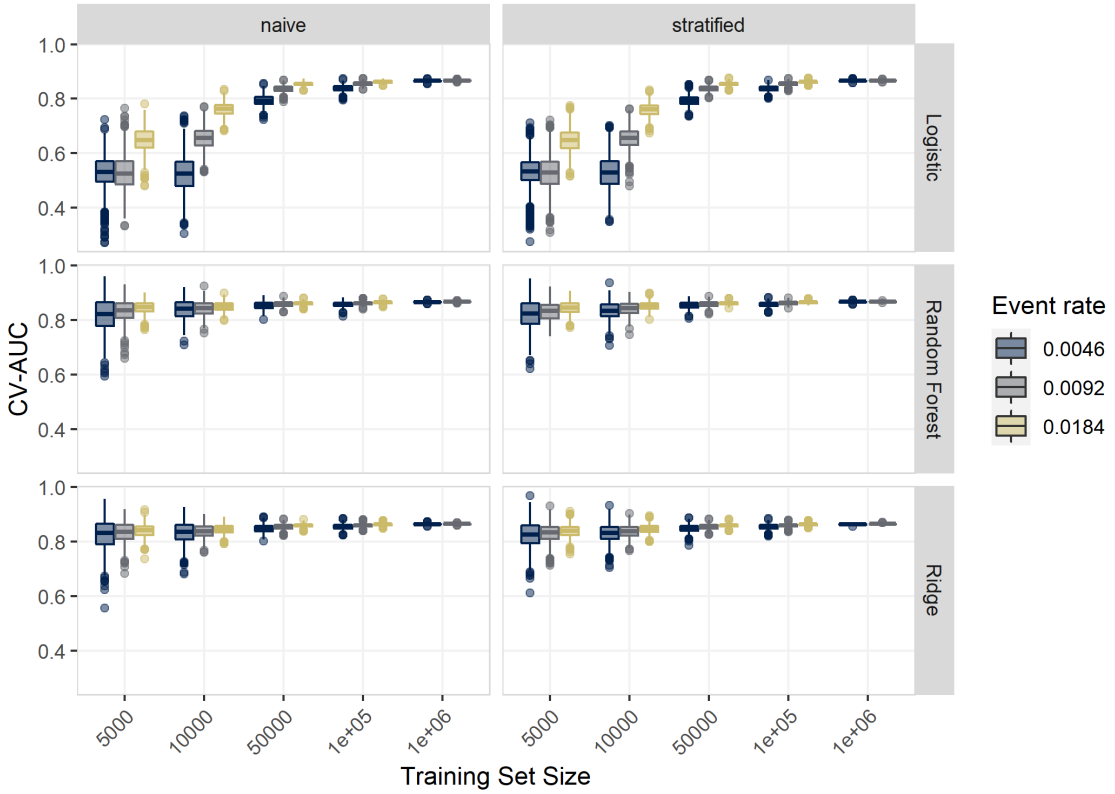

Figure 4.8: Boxplots of iteration-specific cross-validated AUC, stratified by event rate, versus training set size. Colors correspond to event rate. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

Algorithm	SD(AUC_{test})	IQR(AUC_{test})	Range(AUC_{test})	Avg. 95% CI width
Logistic	0.0017	0.0022	0.0118	0.0079
Random Forest	0.0015	0.0021	0.0097	0.0076
Ridge	0.0017	0.0024	0.0123	0.0080

Table 4.4: Measures of spread of test set AUC from one fixed model (trained on a training set of size 1 million) and 1,000 independently sampled test sets of size 1 million, along with the average width of 95% bootstrap percentile-type CI—from sampling across the holdout fold predicted probabilities—with a training set set size of 1 million and a naive sampling strategy. Event rate was $R_0 = 0.0092$.

number of events in the training set increased. However, the variability in test set AUC was less than the variability in cross-validated AUC. At the three smaller training set sizes, the variability for logistic regression was greater than the variability for the ridge and random forest algorthms. The variability in test set AUC was comparable for random forest and ridge regression across training set sizes and event rates. The average test set AUC increased as the number of events in the training set increased. For logistic regression, in addition to the large amount of variability present at $n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$, the average test set AUC at these training set sizes was low—less than 0.8 at all event rates, and less than 0.6 for the two settings with the smallest number of events in the training set. For random forest and ridge regression, the increase in average test set AUC was small.

4.5 Variability Due to Sampling of the Test Set

Sampling the test set from the full dataset introduces sampling variability to the iteration-specific test set AUC value. We investigated this variability by, for each algorithm, training a model on a training set of size one million, then calculating the test set AUC for this model from 1,000 independently sampled training sets of size one million (Section 3.8). Table 4.4 gives the standard deviation, IQR, and range of these test set AUC values, as well as the average width of bootstrap percentile-type confidence intervals at a training set size of 1 million. While the variability of AUC across test sets was small (SD < 0.002 and range < 0.015 for all algorithms), this variability was not small relative to the average width of the confidence intervals.

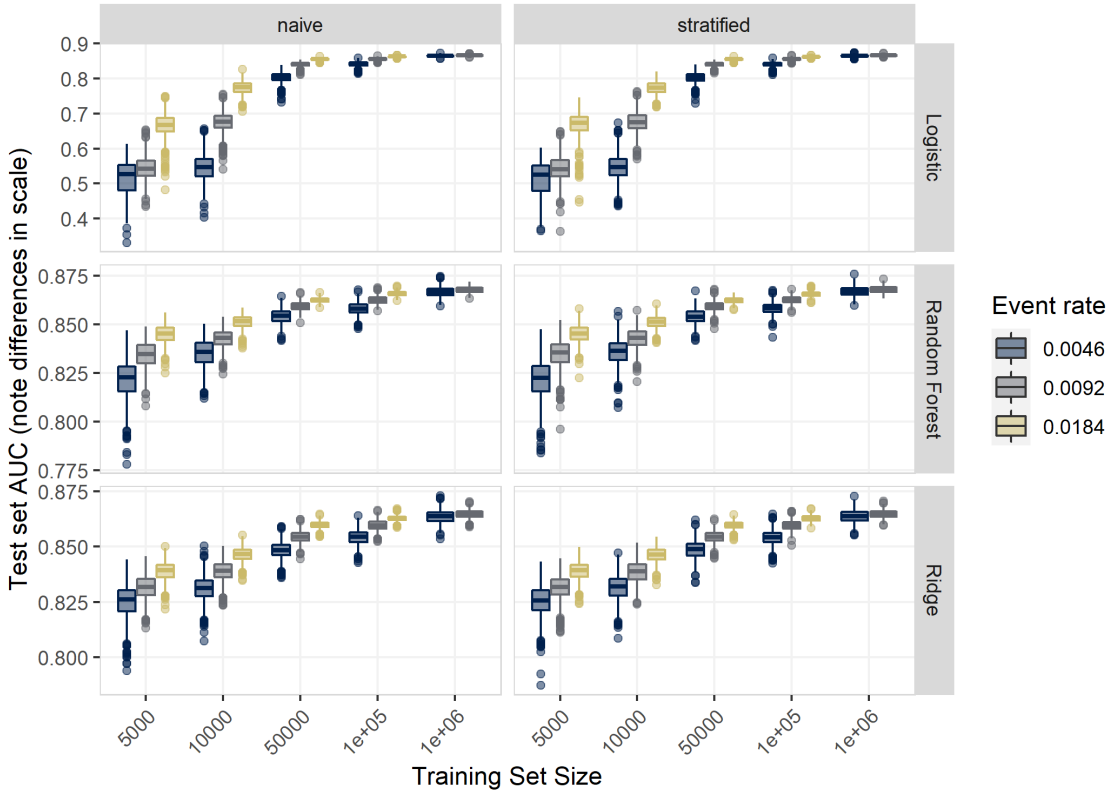

Figure 4.9: Boxplots of iteration-specific test set AUC, stratified by event rate, versus training set size. Colors correspond to event rate. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

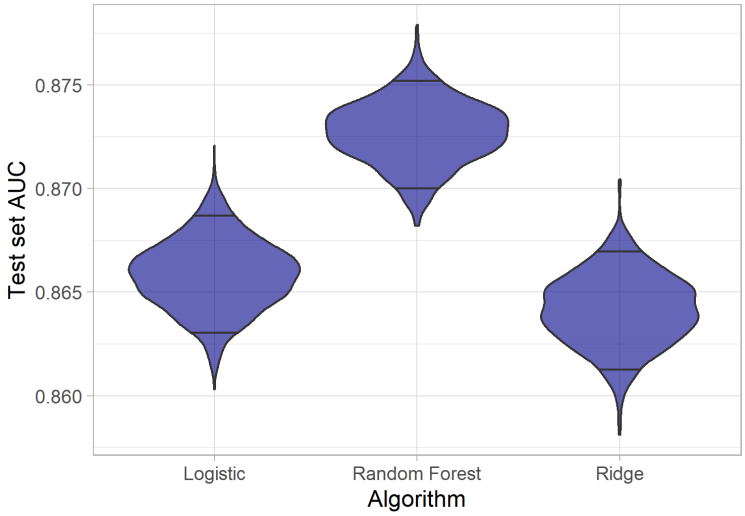

Figure 4.10: Variability of test set AUC for a fixed model across 1,000 independently sampled test sets of size 1 million. The fixed models, one for each algorithm, were trained on a training set of size 1 million. Event rate was $R = 0.0092$. Horizontal lines within violin plots represent the 5th and 95th quantiles.

Accounting for the variability of AUC across test sets in calculating the empirical coverage of resulted in, for the larger training set sizes, empirical coverage closer to the nominal 95% level (Figure 4.11). The difference in coverage increased as the number of events in the training set increase, and was largest at a training set size of 1 million. When the number of training set events was small, there was little difference in coverage. We accounted for this variability by looking at the coverage of the iteration-specific test set AUC ± 0.0016. That is, for each simulation replicate and confidence interval type, we determined if the 95% confidence interval contained AUC_{test}, $\text{AUC}_{\text{test}} + 0.0016$, or $\text{AUC}_{\text{test}} - 0.0016$, all of which were treated the same. The value of 0.0016 was chosen as the average standard deviation of test set AUC (from the same model but independently sampled test sets) across algorithms, rounded to 4 digits.

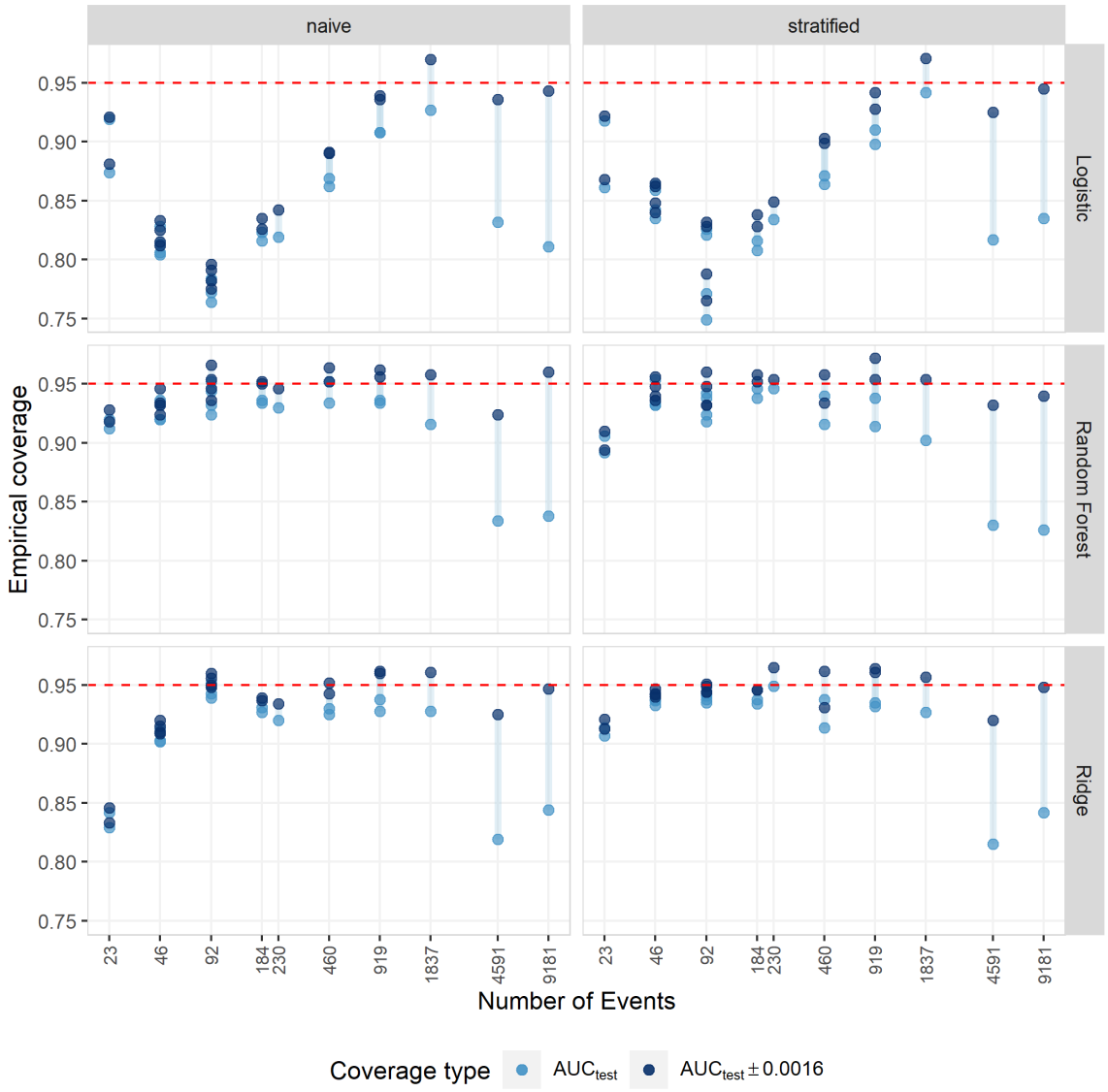

Figure 4.11: Empirical coverage of 95% percentile-type bootstrap confidence intervals, with bootstrap re-sampling done across the aggregated hold out fold predicted probabilities, before and after adjusting for variability due to test set sampling. We adjusted for this variability by calculating the coverage of the iteration-specific test set AUC ± 0.0016, the average (across algorithms) standard deviation of the test set AUC from 1,000 independently test sets with a model trained on a training set size of 1 million and event rate $R_0 = 0.0092$. Colors represent coverage before (lighter blue) or after (dark blue) adjustment. The gray lines represent the distance between the adjusted and unadjusted coverage for a given algorithm, cross-validation sampling strategy, and number of training set events. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

4.6 Ten-by-ten Cross Validation

4.6.1 Ten rounds of cross validation moderately improves performance at small training set sizes

For the two smallest training set sizes, we used ten rounds of ten-fold cross validation (ten-by-ten CV) to estimate prediction performance metrics and to obtain out of sample (holdout fold) predicted probabilities, in addition to using only one round of ten-fold cross validation (one-by-ten CV). The aim of doing both ten-by-ten cross validation and one-by-ten cross validation was to compare the reliability of prediction performance metrics obtained from ten-by-ten cross validation to the reliability under one-by-ten cross validation.

Ten-by-ten cross-validation did not result in meaningfully smaller absolute empirical bias for AUC compared to using only one round of ten-fold cross validation to estimate AUC, except for random forest and ridge regression at the smallest training set size ($n_{\text{train}} = 5,000$) and event rate ($R_0 = 0.0046$). Overall, the absolute empirical bias was not consistently lower with ten-by-ten cross validation than with one-by-ten cross validation (Figure 4.12). The empirical MSE for the ten-by-ten cross-validated AUC was, however, systematically lower than the MSE for one-by-ten cross-validated AUC. The difference between the MSEs was largest when the number of training set events was smallest and was greater under a naive sampling strategy compared to a stratified sampling strategy. However, with slightly more training set events the difference in MSE was very small (Figure 4.13). Similarly, empirical coverage of the test set AUC by bootstrap intervals was higher when bootstrapping was done with the holdout fold predicted probabilities from ten rounds of cross validation, as opposed to one round. However, the difference in coverage was small and decreased as the number of events in the training set increased. The difference in coverage was especially small for ridge regression, and largest for logistic regression (Figure 4.14).

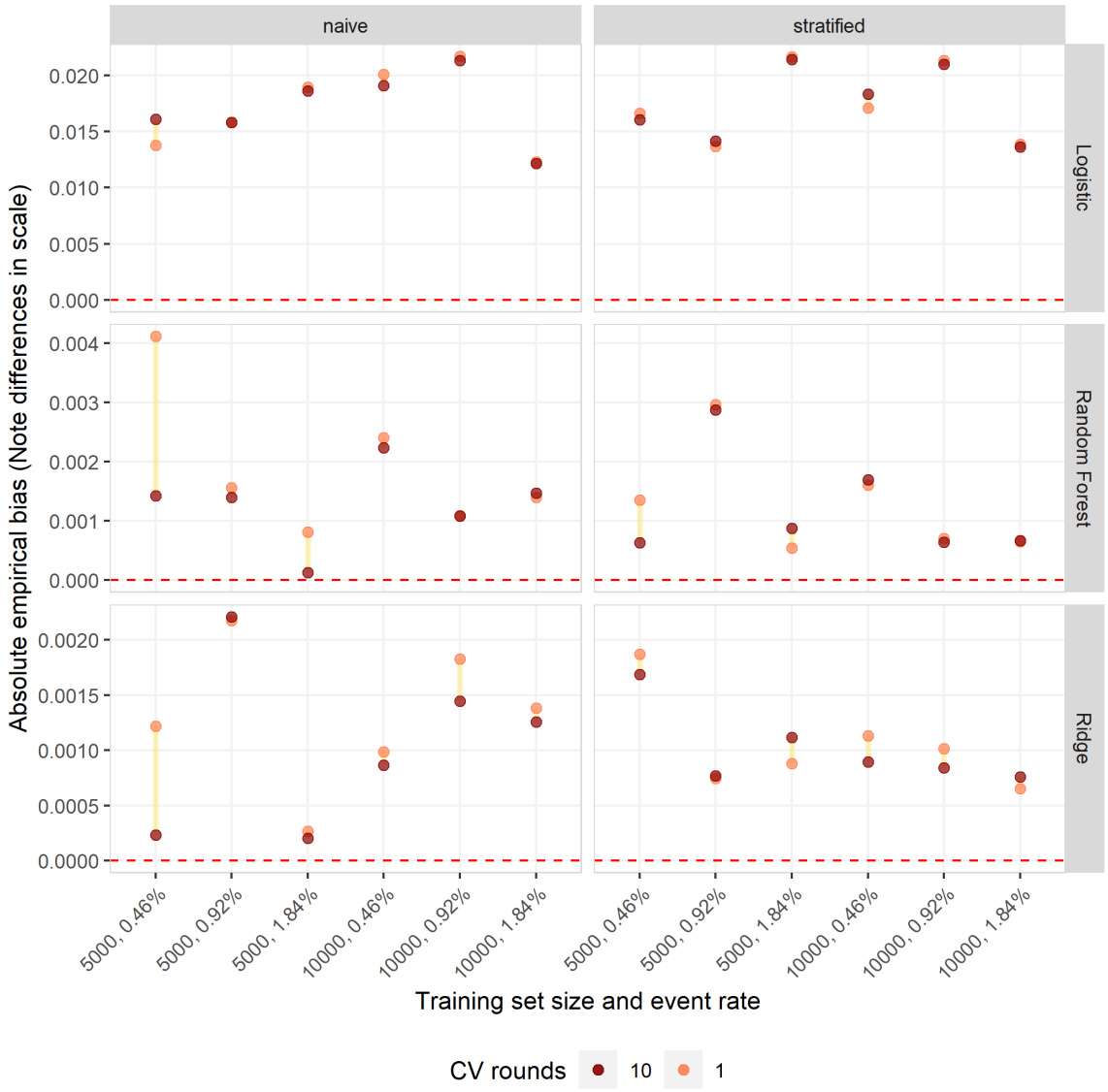

Figure 4.12: Absolute empirical bias for AUC at the two smallest training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$) and all three event rates. Colors represent the number of cross-validation rounds used to obtain cross-validated AUC, ten rounds (dark red) and one round (coral). The yellow lines represent the distance between the bias from 10-by-10 cross-validated AUC and 1-by-10 cross-validated AUC. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

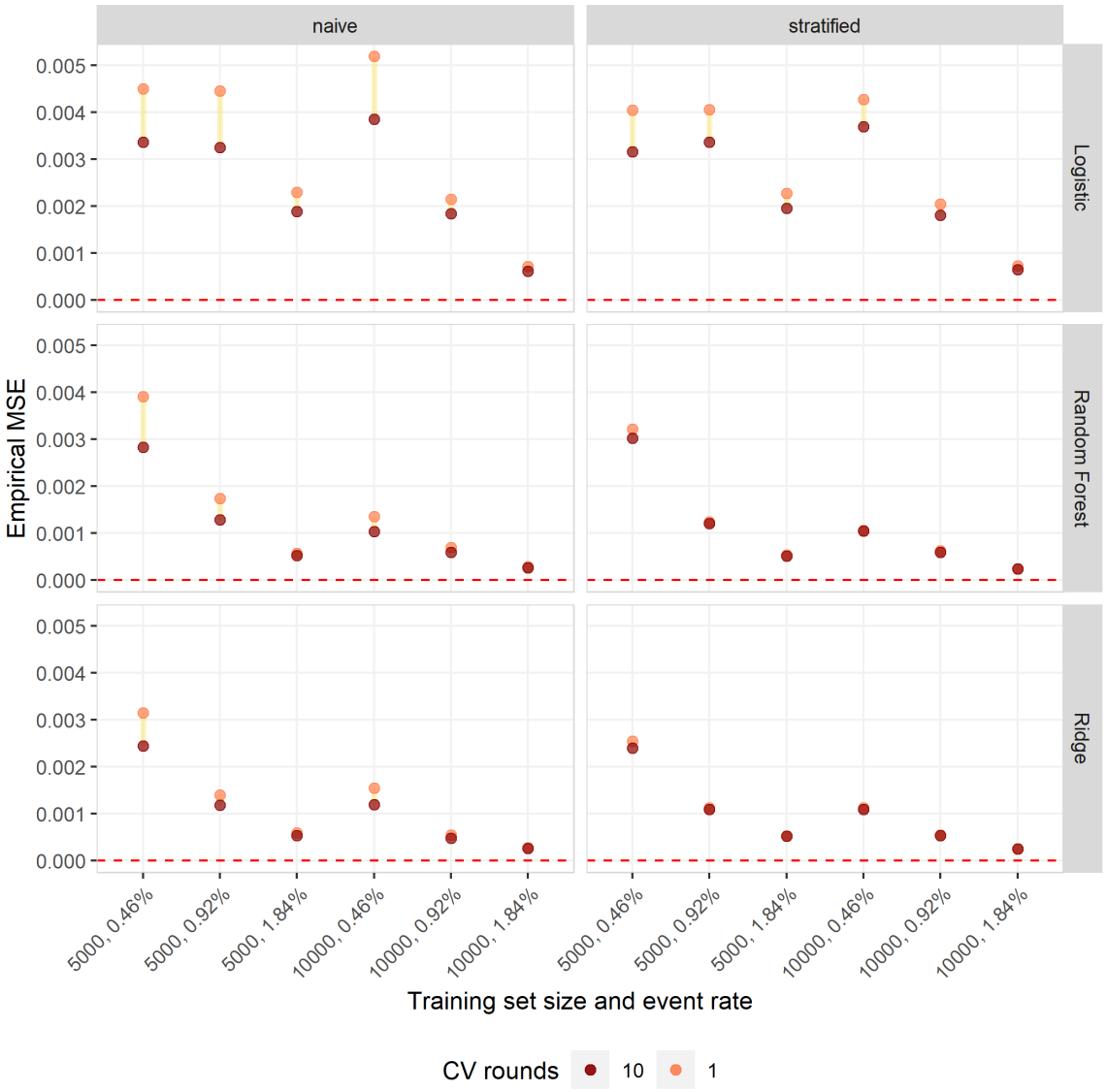

Figure 4.13: Empirical mean squared error for AUC at the two smallest training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$) and all three event rates. Colors represent the number of cross-validation rounds used to obtain cross-validated AUC, ten rounds (dark red) and one round (coral). The yellow lines represent the distance between the MSE from 10-by-10 cross-validated AUC and 1-by-10 cross-validated AUC. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

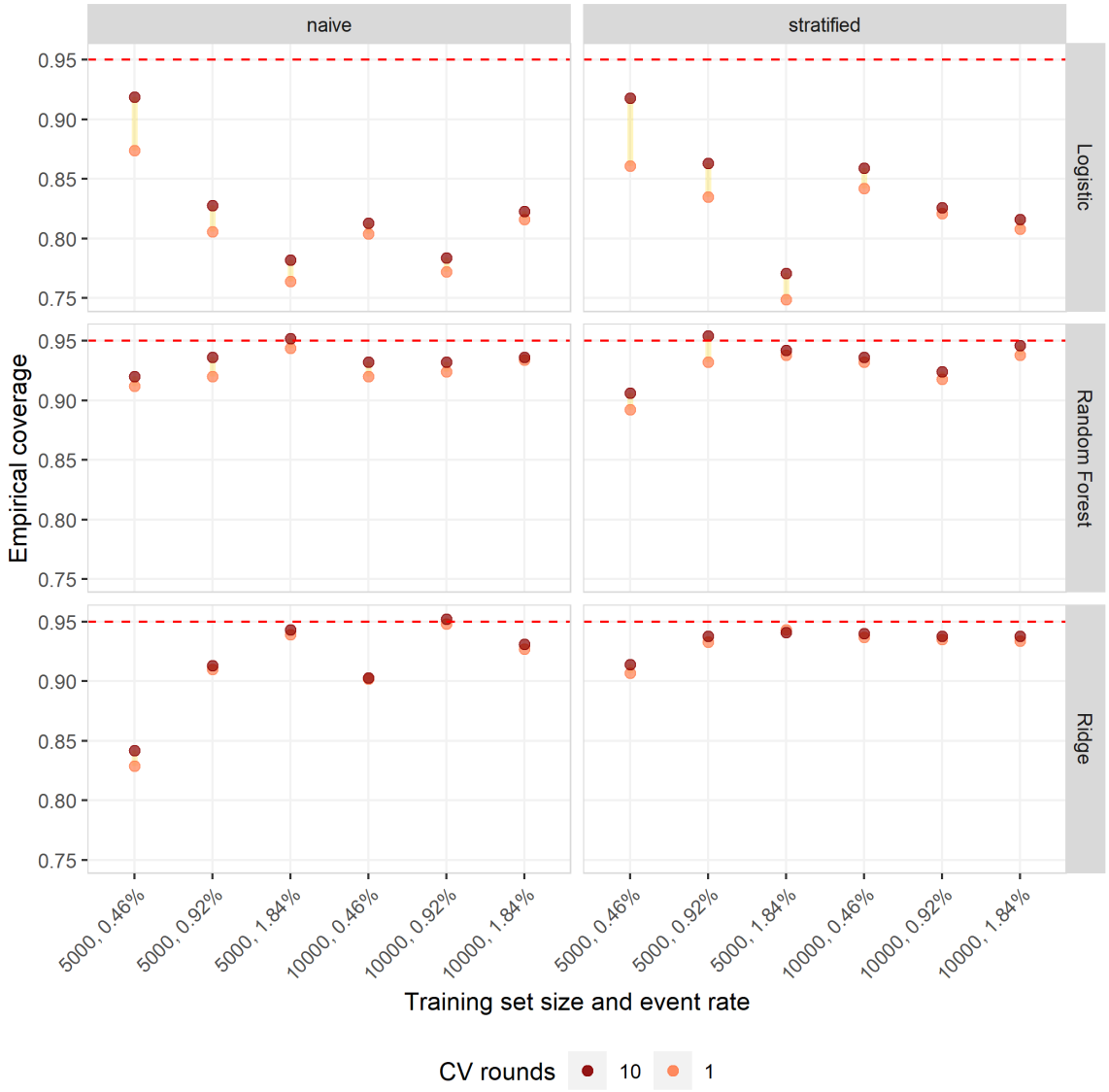

Figure 4.14: Empirical coverage of percentile-type bootstrap intervals, with bootstrap resampling done across the aggregated hold out fold predicted probabilities, at at the two smallest training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$) and all three event rates. Colors represent the number of cross-validation rounds used, ten rounds (dark red) and one round (coral). Yellow lines represent the distance between the coverage using 10-by-10 cross-validation and the coverage using A 1-by-10 cross-validation. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

Chapter 5

RESULTS: OTHER PERFORMANCE METRICS

While AUC was the primary prediction performance metric of interest in this simulation study, we also calculated accuracy, Brier score, F_1 score, $F_{0.5}$ score, NPV, PPV, sensitivity, and specificity using (where applicable) the 90th, 95th, and 99th percentiles of training set predicted probabilities for obtaining predicted events. Like AUC, the reliability of sensitivity —as measured by empirical bias, empirical MSE, empirical variability of the cross-validated estimate across simulation replicates, and the empirical coverage of the 95% CIs—depended on the number of events in the training set but not directly on event rate. That is, there was not a dependence on event rate beyond, for a given training set size, there being more events with a larger event rate. The reliability of specificity also had little dependence on event rate, although empirical MSE and the variability of cross-validated specificity estimates across simulation replicates had greater dependence on the size of the training set than on the number of events in the training set. The reliability of the other prediction performance metrics, however, did depend directly on rate.

In the following sections we report the simulation results for each metric in turn. We report the results for accuracy, sensitivity, and specificity—all at the 95th percentile threshold—in greater detail. For the other metrics, we note how the patterns in reliability with event rate, training set size, sampling strategy, and number of training set events are similar to or different from those of AUC, accuracy, and sensitivity. We also note any other observations of interest. Table 2.1 gives an overview of the patterns observed with these prediction performance metrics, while Appendix B and Appendix C contain tables of the measures of performance organized by metric, algorithm, and percentile threshold.

As with AUC, for many performance metrics—all except Brier score—the empirical coverage of the 95% bootstrap confidence intervals dropped for a training set size of 1 million, although the amount by which the coverage dropped varied by metric and percentile thresh-

old of predicted probabilities used define a predicted event. Unlike with AUC, however, we did not formally investigate the variability of the iteration-specific test set value for these other performance metrics of interest, although this variability surely exists and similarly would impact the empirical coverage of the boostrap intervals. This is noted here to provide context to coverage results for these other prediction performance metrics.

Metric	Drivers of trends in bias, MSE, and Var(CV-value)*	Bias (pos. or neg., on average)	Trends in test set value with rate
Accuracy	rate, n_{train}	negative	lower with smaller rate
Brier Score	rate, n_{train}	positive	lower with smaller rate
F_1 Score	rate, n_{train}	negative	lower with smaller rate
$F_{0.5}$ Score	rate, n_{train}	positive for $n_{\text{train}} = 5,000$ and $10,000$, then negative	lower with smaller rate
NPV	rate, n_{train}	negative	higher with smaller rate
PPV	rate, n_{train}	negative	lower with smaller rate
Sensitivity	n_{event} no direct dependence on rate	negative	no difference with rate†
Specificity	$n_{\text{non-event}}$ no direct dependence on rate	negative	no difference with rate‡

* Absolute bias, MSE and Var(CV-value) decreased as the training set size increased
† Expect for logistic regression at the two smallest training set sizes
‡ Except for logistic regression and the smallest training set size

Table 5.1: Patterns in empirical bias, empirical MSE, variance of the cross-validated estimates, and average test set value for prediction performance metrics other than AUC.

5.1 Accuracy

Coverage of the 95% bootstrap confidence intervals (Figure 5.1, percentile-type confidence intervals with bootstrap re-sampling done across the aggregated hold out fold predicted probabilities) for accuracy at the 95th percentile threshold depended on the algorithm and training set size. For logistic regression, the lowest empirical coverage occurred at the two smallest number of training set events, $n_{\text{event}} = 23$ and $n_{\text{event}} = 46$. At a training set size of 1 million coverage was approximately 80%. Otherwise, coverage of the confidence

intervals was near 95%. With logistic regression, there was little difference in coverage between a naive and stratified sampling strategy for creating cross-validation folds. For ridge regression, empirical coverage was near the nominal 95% value for all but a training set size of 1 million and for the smallest number of events in the the training set with a naive sampling strategy. In all combinations of training set size, event rate, and sampling strategy, empirical coverage for ridge regression was greater than 80%. For random forest, empirical coverage at a training set size of 1 million was only approximately 40%, while coverage at other training set sizes was at or greater than 80%. For random forest, coverage appeared to be dependent on the size of the training set, which was not the case for logistic regression and ridge regression. While Figure 5.1 only shows coverage of one type of bootstrap confidence interval, the pattern and levels of empirical coverage were similar for the other types of confidence intervals. Additionally, on average, coverage was slightly higher at a higher percentile threshold.

Empirical bias for accuracy was generally negative, and the absolute empirical bias decreased as the training set size increased (Figure 5.2). The magnitude of the empirical bias for random forest and ridge regression was similar. The absolute empirical bias for logistic regression was larger at the smaller training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$), however the magnitude of the bias for all three algorithms was similar at the larger training set sizes. At the smallest training set size and event rate a naive sampling strategy resulted in greater absolute bias compared to a stratified sampling strategy. The difference was most pronounced for ridge regression and random forest. However, at larger training set sizes there was little to no difference in empirical bias between a naive and stratified sampling strategy.

The MSE for accuracy depended on training set size and event rate (Figure 5.3). In general, MSE was lower at larger training set sizes and was higher for larger event rates. The MSE for random forest and ridge regression was similar at all training set sizes. The empirical MSE for logistic regression was much larger than that for random forest and ridge regression at the smallest training set size. Except at the smallest training set size and event rate, there was little or no difference in empirical MSE between a naive and stratified sampling strategy.

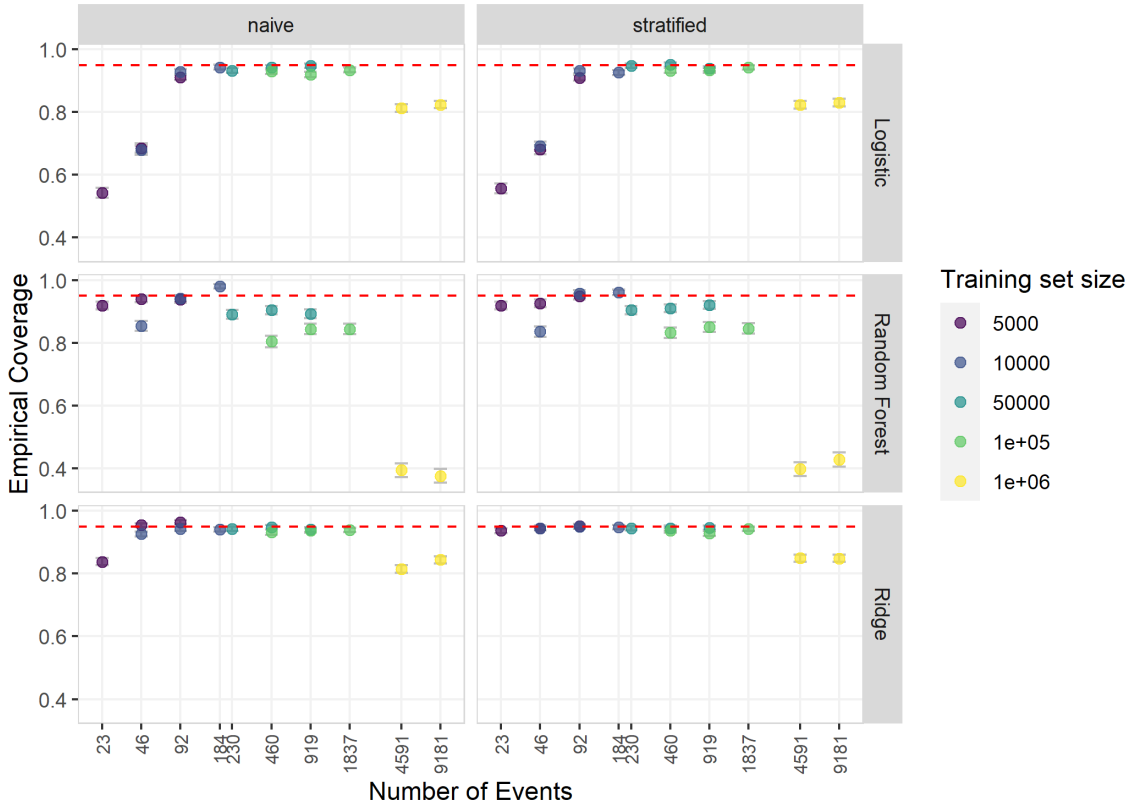

Figure 5.1: Empirical coverage of test set accuracy at the 95th percentile, using percentile-type bootstrap confidence intervals, versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size, while the red dotted lines represent the nominal 95% coverage level. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

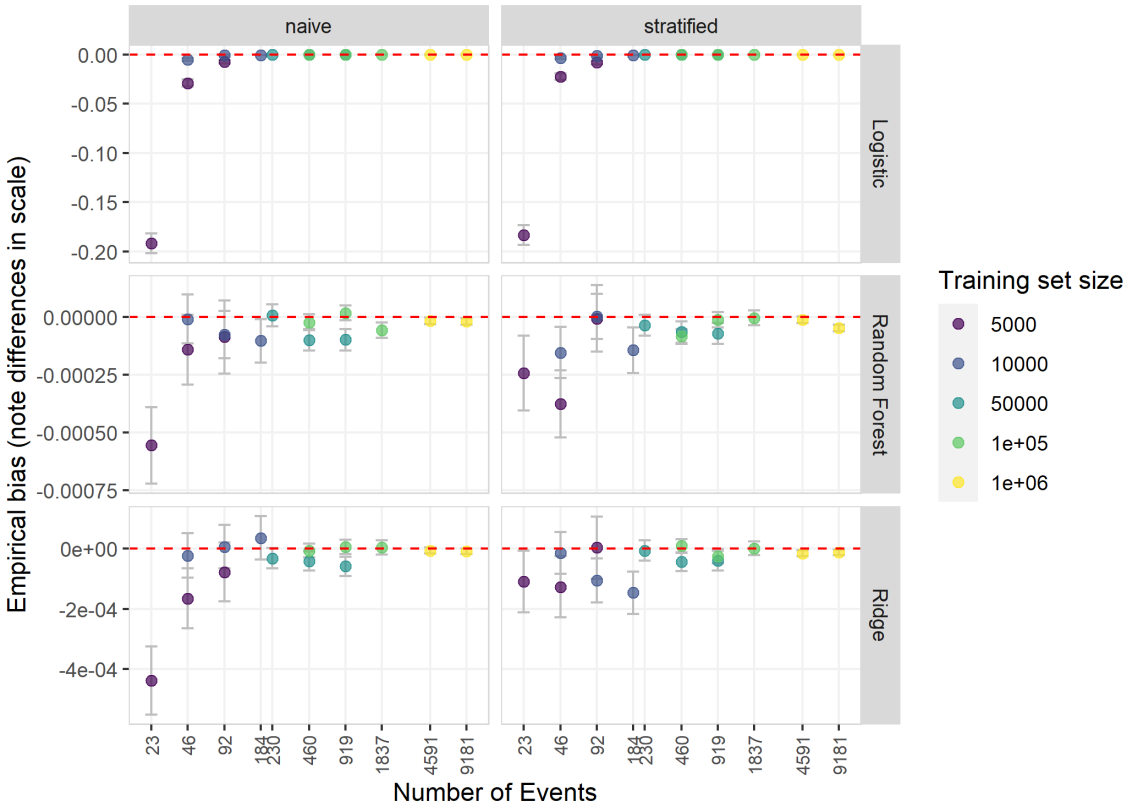

Figure 5.2: Empirical bias for accuracy at the 95th percentile versus the number of events in the training set. Positive bias indicates that, averaging across iterations, the cross-validated accuracy was higher than the test set accuracy. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

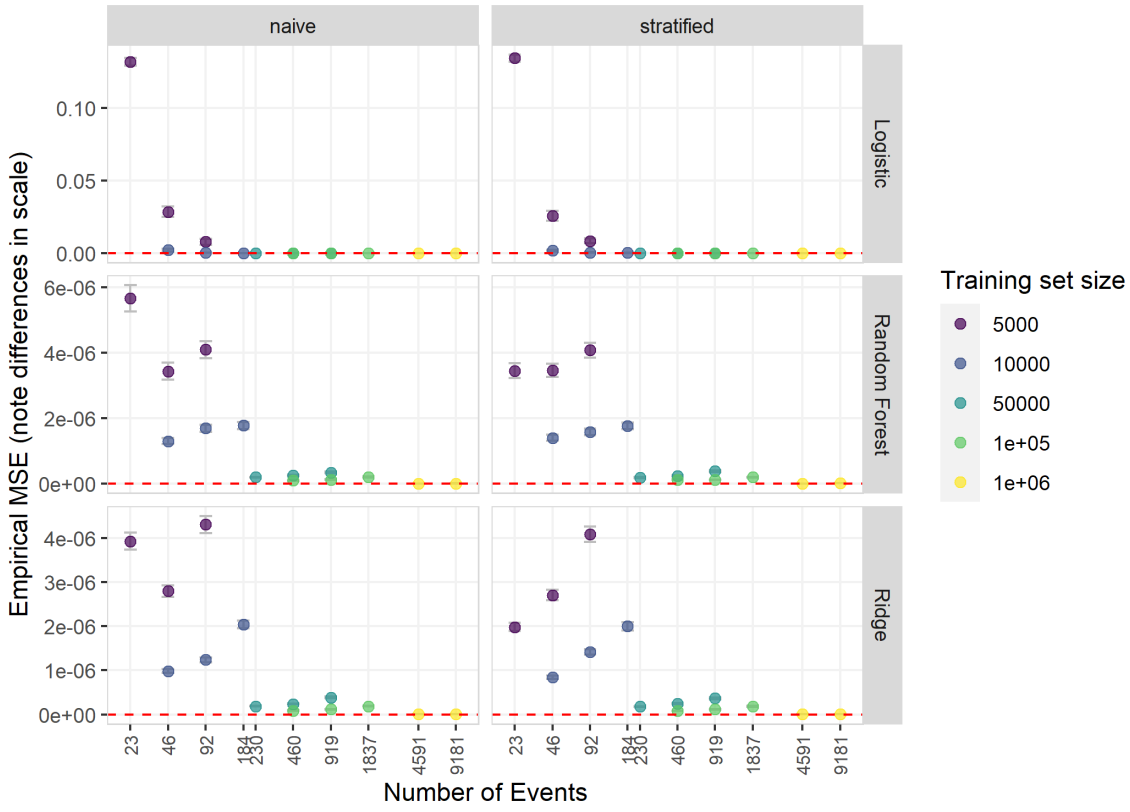

Figure 5.3: Empirical mean squared error for accuracy at the 95th percentile versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

Like MSE, the variability of the cross-validated accuracy across iterations depended on the training set size and event rate. The variance of CV-accuracy was lower at larger training set sizes and higher for larger event rates (Figure 5.4). The variance of CV-accuracy was, at the smaller training set sizes, larger for logistic regression than for random forest or ridge regression. The cross-validated accuracy estimate was, for some iterations with the logistic regression model, very low at the smallest training set size. For example, at $n_{\text{train}} = 5,000$ and $R = 0.0046$ and with a stratified sampling strategy, the minimum 95th percentile CV-accuracy observed was only 0.0046. Comparatively, for random forest and ridge regression the equivalent minimum cross-validated accuracies were 0.939 and 0.944, respectively. At the larger training set sizes the three algorithms had similar variances and similar values for cross validated accuracy.

5.2 Sensitivity

The empirical coverage for sensitivity at the 95th percentile was driven by number of events in the training set (Figure 5.5). For all algorithms and cross-validation fold sampling strategies, coverage was lowest for a training set of size 1 million. There was little difference in empirical coverage between a naive and stratified sampling strategy, except at the smallest number of training set events and, for random forest only, at a training set size of 1 million. In these instances, a naive sampling strategy resulted in worse coverage. Overall, the best coverage was observed for ridge regression, with empirical coverage slightly below the nominal 95% value in almost all instances. Coverage for random forest was similar to that for logistic regression, except when the number of events in the training set was large. As the number of events in the training set increased beyond 500, the empirical coverage for random forest decreased. In contrast, for logistic regression and ridge regression there was only a decrease in coverage at the two largest number of training set events. Coverage for logistic regression first decreased between $n_{\text{event}} = 23$ and $n_{\text{event}} = 46$, then increased as the number of events increased until dropping for two largest number of training set events. Except at the smallest training set sizes, there was little difference in coverage between the types of bootstrap confidence intervals. Additionally, there was slightly higher coverage on average for test set sensitivity at a lower percentile threshold. For example, coverage

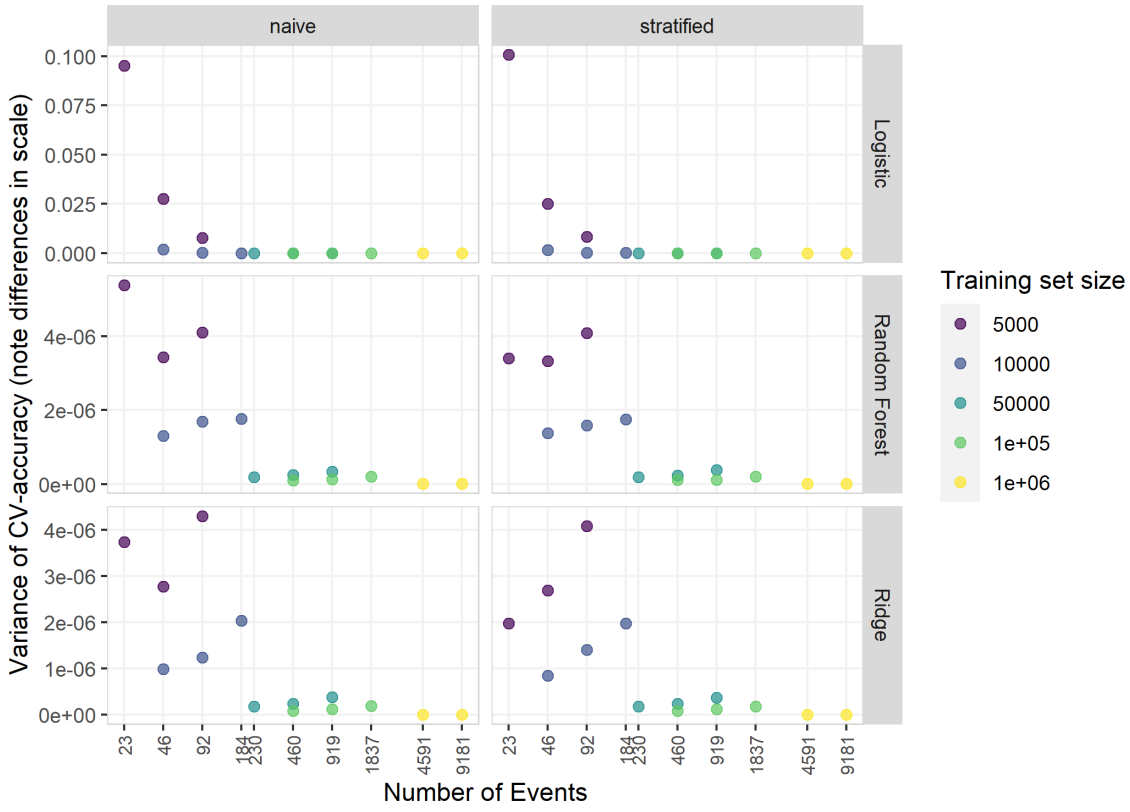

Figure 5.4: Variance of cross-validated accuracy at the 95th versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

was slightly higher for sensitivity at the 90th percentile compared to sensitivity at the 95th percentile.

As with AUC, the empirical bias (Figure 5.6), empirical MSE (Figure 5.7), and variance of the CV sensitivity estimate across simulation replicates (not shown) trended toward zero with the number of events in the training set. The random forest and ridge regression algorithms had similar values of bias, MSE, and variance. The bias, MSE, and variance for logistic regression were notably larger—compared to those for ridge regression and random forest— when the number of events in the training set was small. However, the values became more similar between all three algorithms as the number of events increased. Bias was mostly negative, indicating that, on average, the cross-validated sensitivity underestimated the test set sensitivity. Additionally, bias, MSE, and variance were similar at the 90th, 95th, and 99th percentile thresholds.

5.3 Brier Score

Variance of the cross-validated Brier score estimate, empirical MSE, and empirical bias all depended on the event rate in addition to the training set size. For a given training set size, variance, bias, and MSE were higher at a larger event rate. As the training set size increased, the absolute difference in empirical bias, MSE and variance between the event rates decreased. The average test set Brier score also depended on the event rate, with a higher event rate corresponding to a higher Brier score. For ridge regression and random forest, empirical coverage of the 95% bootstrap confidence intervals was 100% for all types of bootstrap intervals, except for at the smallest event rate and smallest training set size. For logistic regression, empirical coverage of the bootstrap confidence intervals was 100% for all but the two smallest training set sizes, $n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$.

5.4 F_1 Score

The variance of the cross-validated F_1 score, its empirical MSE, and its empirical bias depended on event rate in addition to the training set size. At the 90th and 95th percentile thresholds, for a given training set size the variance and MSE were higher at a larger event rate, with the absolute difference between event rates decreasing as the training set size

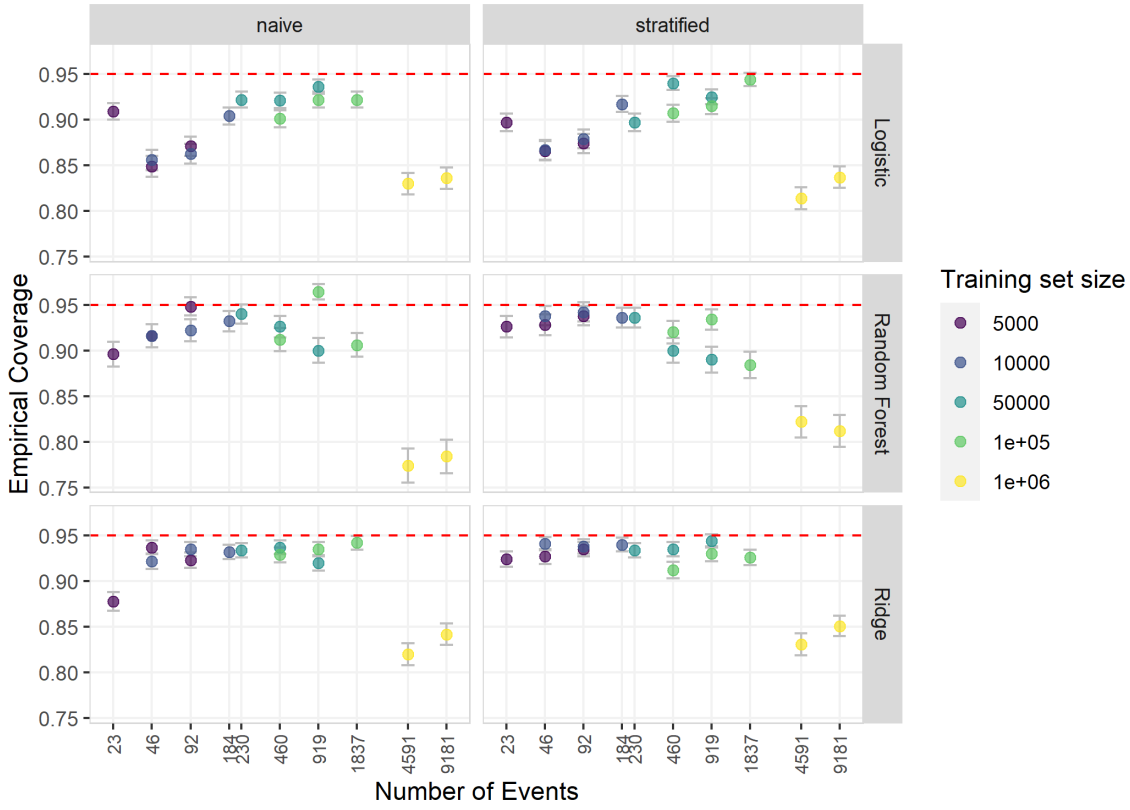

Figure 5.5: Empirical coverage of test set sensitivity at the 95th percentile, using percentile-type bootstrap confidence intervals, versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size, while the red dotted lines represent the nominal 95% coverage level. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

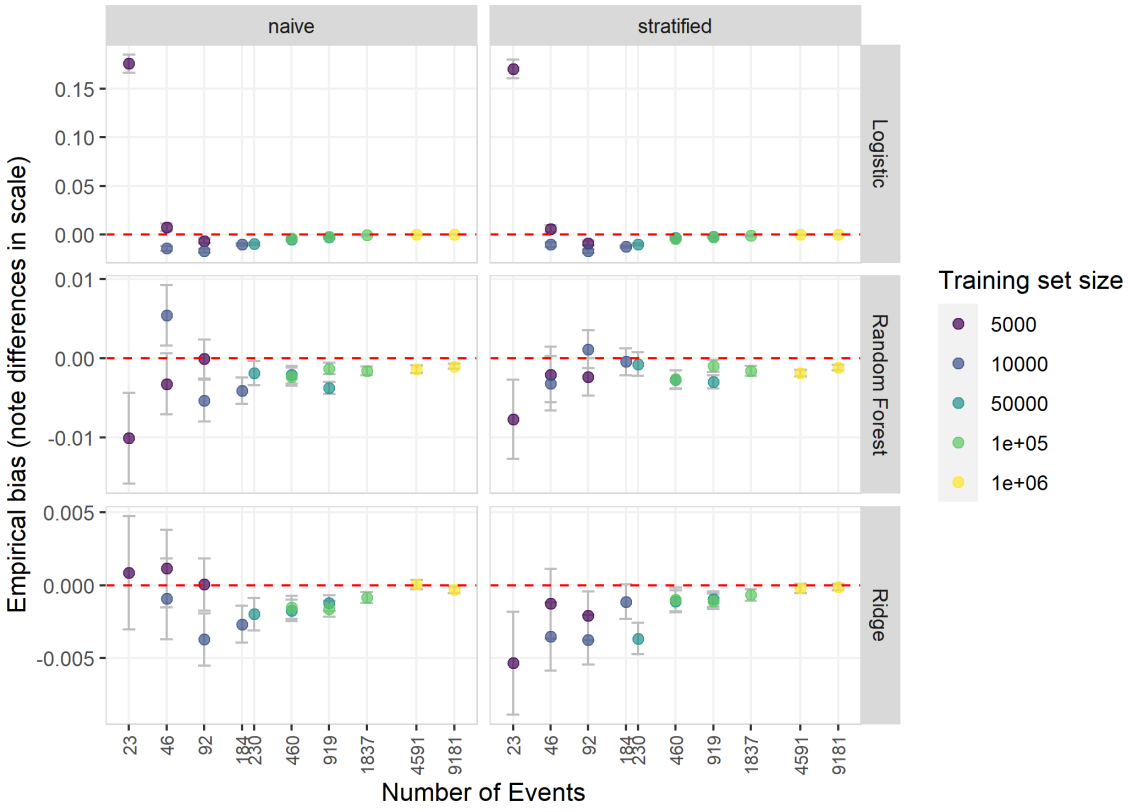

Figure 5.6: Empirical bias for sensitivity at the 95th percentile versus the number of events in the training set. Positive bias indicates that, averaging across iterations, the cross-validated sensitivity was higher than the test set sensitivity. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

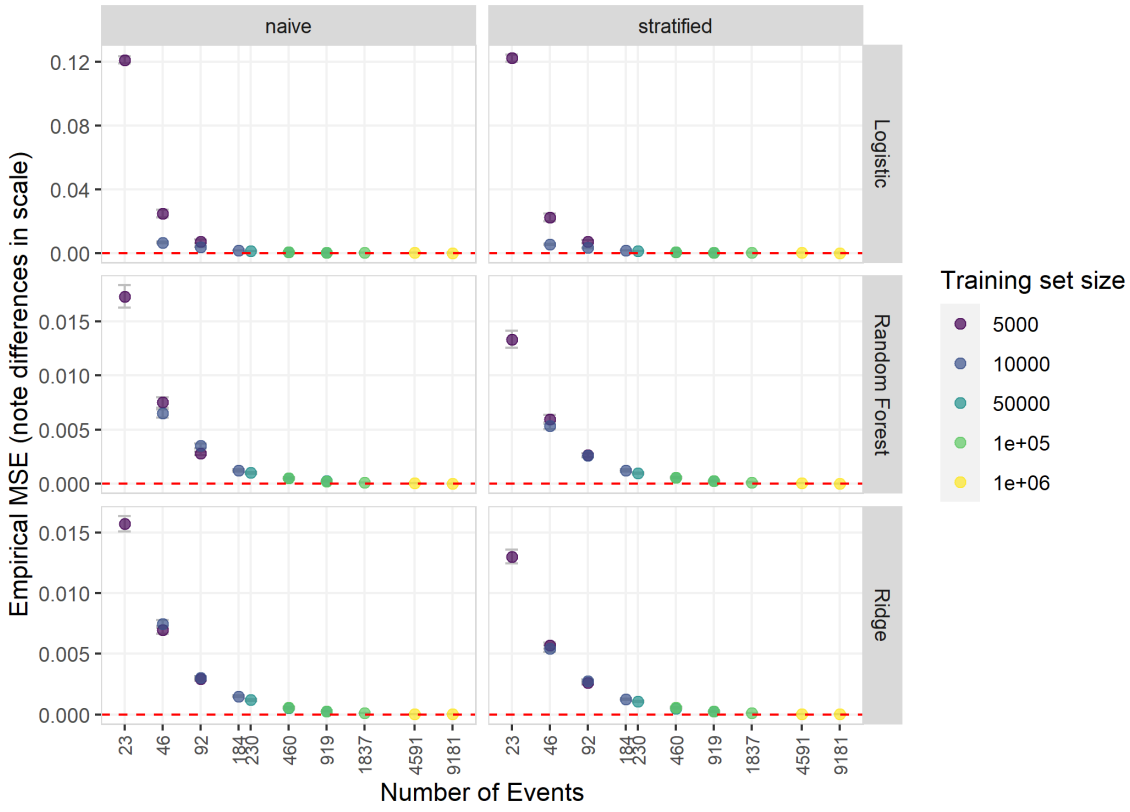

Figure 5.7: Empirical mean squared error for sensitivity at the 95th percentile versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

increased. Variance and MSE at the 99th percentile depended more on training set size than event rate. At this 99th perecentile threshold, the MSE and variance for the intermediate event rate was generally the largest and the MSE and variance for the largest event rate the smallest. The pattern in empirical bias with event rate was less clearly defined than for variance and MSE. However, the absolute empirical bias decreased as the training set size increased. At the smaller training set sizes and with random forest and ridge regression, a stratified sampling strategy also resulted in lower absolute empirical bias than a naive sampling strategy.

For the random forest and ridge regression algorithms, coverage of the 95% bootstrap confidence intervals formed an inverted U-shape, with lower empirical coverage at $n_{\text{train}} = 5,000$ and $n_{\text{train}} = 1$ million than at the intermediate training set sizes. Figure 5.8 shows this shape for PPV at the 95th percentile threshold. Coverage was higher at a lower percentile threshold. For example, looking specifically at ridge regression, coverage at the 90th percentile threshold was always above 95% while coverage at the 99th percentile was always below 95%. For logistic regression, coverage also formed an inverted U-shape with the exception of the two simulations (naive and stratified sampling strategy) with the smallest number of training set events, where coverage was higher. Overall, there was little difference in coverage between a naive and stratified sampling strategy.

5.5 $F_{0.5}$ Score

The patterns in $F_{0.5}$ score empirical absolute bias, MSE, coverage, and variance of the cross-validated estimate across simulation replicates were very similar to the patterns for F_1 score. However, the empirical $F_{0.5}$ score tended to be positive at smaller training set sizes ($n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$), then comparatively very near zero for the larger training set sizes.

5.6 Negative Predictive Value

At training set sizes greater than 5,000, the absolute empirical bias, empirical MSE, and variance of the cross-validated NPV estimates across simulation replicates were—at a given training set size—greater at a larger event rate. The absolute difference in bias, MSE, and

variance between event rates decreased as the training set size increased. At a training set size of 5,000 there was greater variability in the performance measures. The magnitudes of bias, MSE, and variance were similar between the three algorithms. There was also little difference in empirical coverage of the 95% bootstrap confidence intervals between algorithms. Coverage was higher for NPV at a higher percentile threshold. However, coverage of all interval types was greater than 95% for all but a training set size of 1 million. For all performance measures, there was little difference between a naive and stratified sampling strategy.

5.7 Positive Predictive Value

As with NPV, the empirical bias, empirical MSE, and variance of the cross-validated PPV estimates depended on the event rate as well as the training set size. With PPV, the empirical bias was noticeably greater in magnitude at a training set size of of 5,000 and 10,000 in comparison to larger training set sizes. This pattern in bias with training set size dominated over any pattern with event rate. However, there was a clear pattern in MSE and variance of the cross-validated PPV estimates with event rate. For a given training set size, the MSE and variance were greater at a larger event rate, and the absolute difference in MSE and variance between event rates decreased as the training set size increased. As with F_1 score, the coverage of the 95% bootstrap confidence intervals formed an inverted U-shape (Figure 5.8). With all performance measures, there was little difference between a naive and stratified sampling strategy for creation of cross-validation folds.

5.8 Specificity

The empirical MSE and empirical bias for specificity, and the variance of cross-validated specificity across simulation replicates, depended primarily on the number of non-events in the training set. However, since all event rates considered in this simulation study were small, $n_{\text{non-event}} = (1 - R) \times n_{\text{train}} \approx n_{\text{train}}$, and thus the biggest differences in reliability measures were between training set sizes, not between different event rates for a given training set size. MSE and variance were similar across event rates for a given training set size, although the MSE and variance were slightly higher at a smaller number of non-

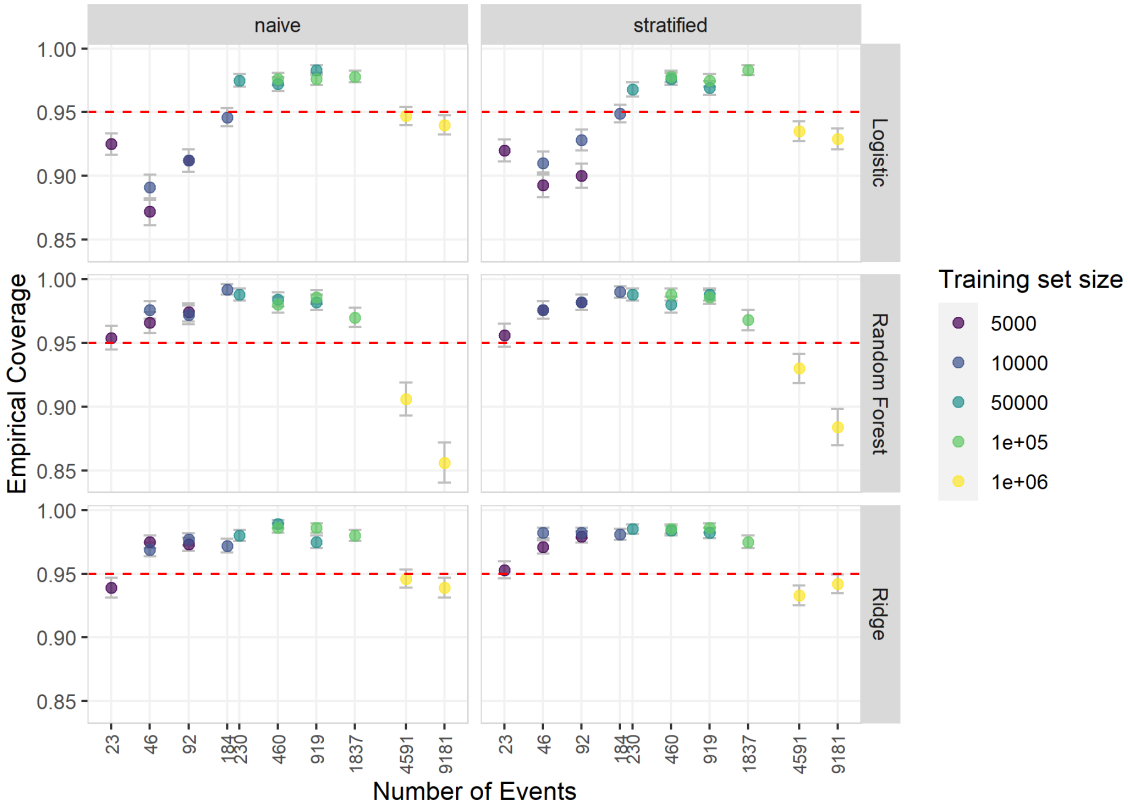

Figure 5.8: Empirical coverage of test set PPV at the 95th percentile, using percentile-type bootstrap confidence intervals, versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size, while the red dotted lines represent the nominal 95% coverage level. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

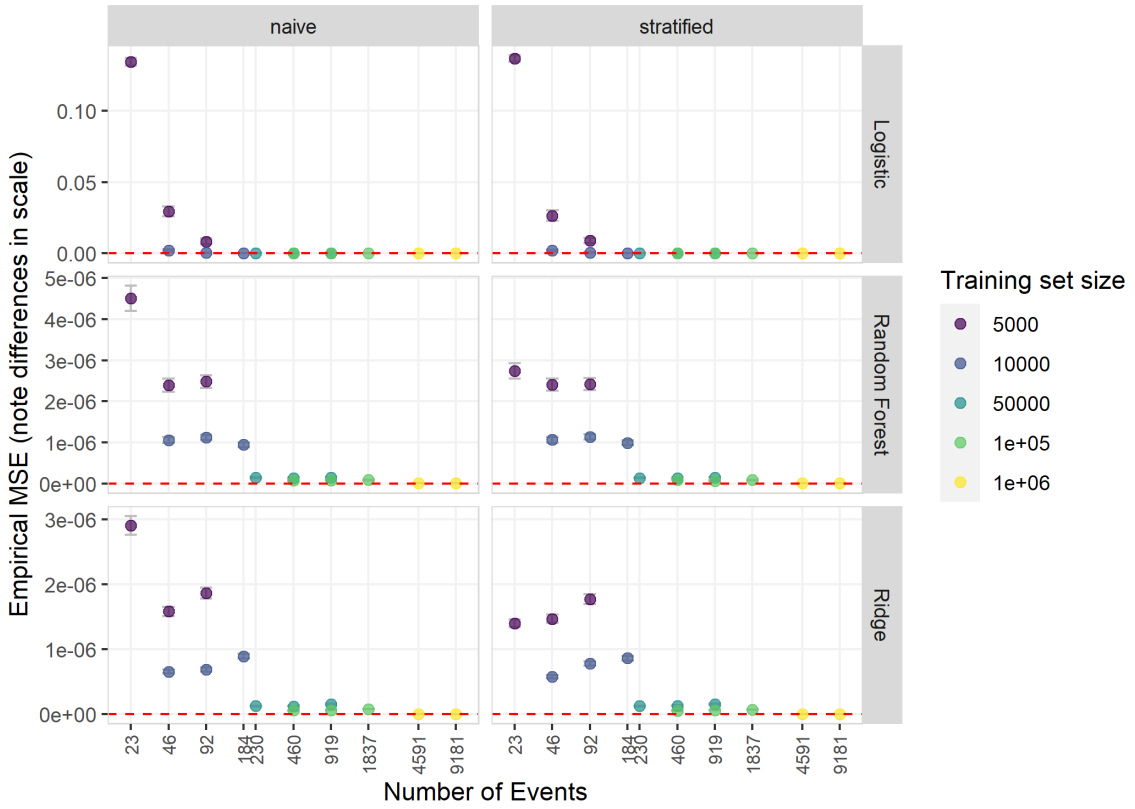

Figure 5.9: Empirical mean squared error for specificity at the 95th percentile versus the number of events in the training set. Colors correspond to the training set size, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

events (larger number of events) for a given training set size (Figure 5.9 for MSE, figure for variance not shown). Figure 5.10 shows the empirical bias for specificity at the 95th percentile threshold plotted against the size of the training set. Especially on the log-scale, in this setting n_{train} is a good approximation of $n_{\text{non-events}}$. As the number of training set non-events increased, the absolute empirical bias decreased. However, there was little difference in empirical bias comparing different rates for a given training set size, after accounting for variability in the empirical bias due to the simulation Monte Carlo error.

There also was some dependence on the number of training set non-events in the empirical coverage of the 95% bootstrap confidence intervals. However, this dependence appeared

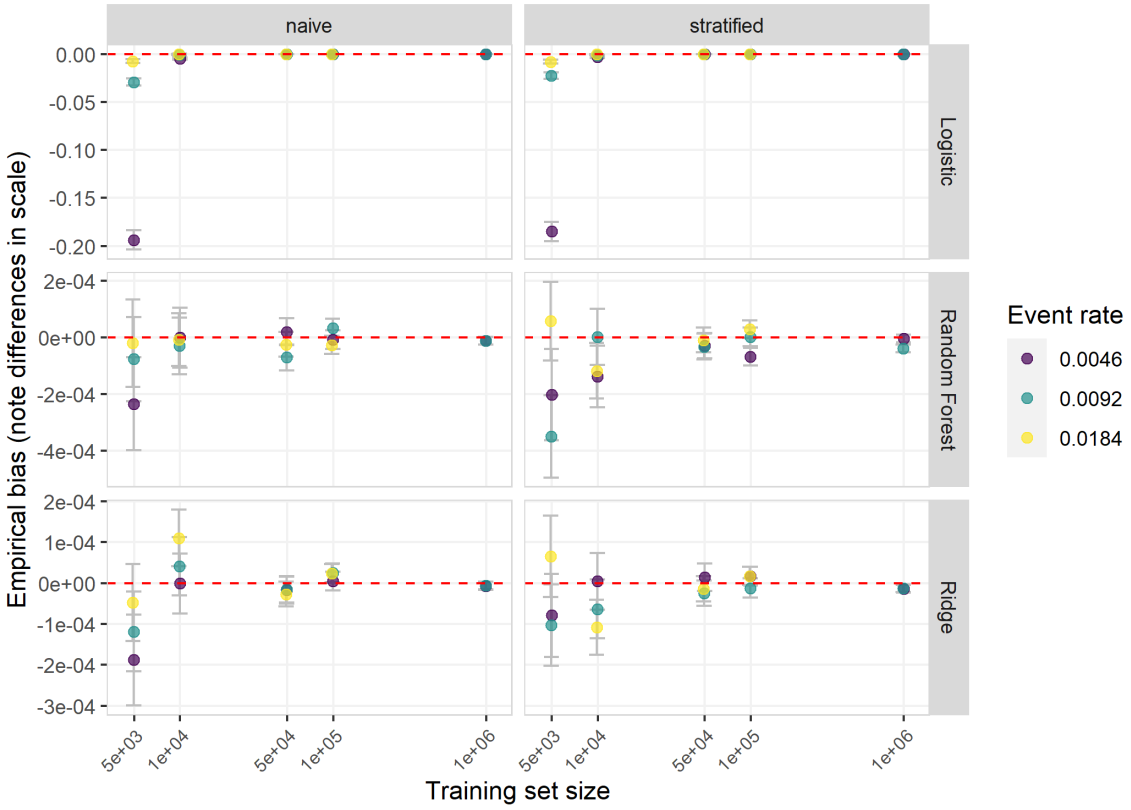

Figure 5.10: Empirical bias for specificity at the 95th percentile versus training set size. Colors correspond to the event rate, and Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom). Note that training set size was used for the x-axis since $\log(n_{\text{non-events}}) \approx \log(n_{\text{train}})$ at the small events rates used.

only for the random forest algorithm, with which there was slightly lower coverage with a larger training set size (which, again, corresponds to a larger number of training set non-events). For ridge regression and logistic regression, this dependence on the number of training set non-event was less apparent. And, for ridge regression, coverage of the bootstrap intervals for specificity at the 90th, 95th, and 99th percentile thresholds was at or near the nominal 95% level for almost all combination of training set size and event rate (Figure 5.11).

There was little difference in mean squared error or variance of the cross-validated estimates between a naive and stratified sampling strategy, except for at the smallest event rate and training set size (Figure 5.9). In this instance, MSE and variance were higher with a naive sampling strategy. There was little difference in bias or coverage comparing naive and stratified sampling strategies at all number of training set sizes and event rates (Figures 5.11, 5.10).

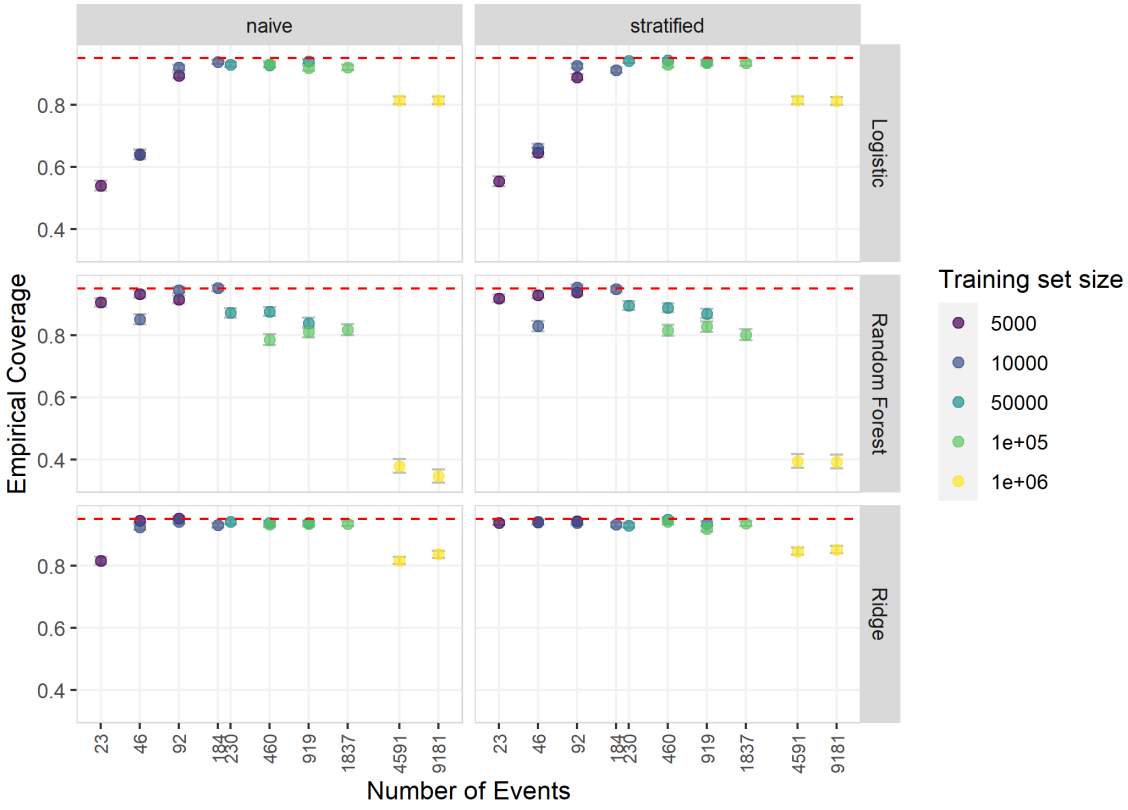

Figure 5.11: Empirical coverage of test set specificity at the 95th percentile, using percentile-type bootstrap confidence intervals, versus number of events in the training set. Bootstrap re-sampling was done across the aggregated holdout fold predicted probabilities. Colors correspond to the training set size, while the red dotted lines represent the nominal 95% coverage level. Monte Carlo error is displayed in error bars. Grid columns correspond to the sampling strategy used for cross-validation, naive (left) or stratified (right). Grid rows correspond to the algorithm used: logistic regression (top), random forest (middle), or ridge regression (bottom).

Chapter 6

DISCUSSION

6.1 Poor performance of logistic regression with small n_{event}

Logistic regression performed particularly poorly when the number of events in the training set was small. Empirical bias, MSE, and confidence interval coverage for AUC were all worse than those for random forest or ridge regression until a training set size of 100,000. With a very small number of events ($n_{\text{event}} \leq 92$), the absolute empirical bias for AUC was approximately ten times greater for logistic regression than for random forest or ridge regression (Figure 4.4). While with the random forest and ridge regression algorithms empirical coverage of most types of bootstrap intervals for the iteration-specific test set AUC was above 90% for all but the smallest number of training set events and $n_{\text{train}} = 1$ million (Tables 4.1, 4.2), with logistic regression empirical coverage of the bootstrap intervals was not greater than or equal to 90% until $n_{\text{event}} = 919$.

We also saw evidence of the poor performance of logistic regression when there were few events in the training set through the values and variability of CV-AUC estimates and test set AUC. Variability of the cross-validated and test set AUC values across simulation replicates was higher for logistic regression than for random forest or ridge regression, and there could be large differences between the cross-validated AUC estimate and the test set AUC (Figures 4.8, 4.9, 4.5). The average test set AUC and average cross-validated AUC for logistic regression were very low when there were few training set events: below 0.6 for the two smallest n_{event}, and below 0.7 for the third smallest number of events in the training set, $n_{\text{event}} = 92$. At the very smallest number of events, $n_{\text{event}} = 23$, in many replicates the logistic regression test set AUC was even below 0.5.

This poor performance of logistic regression when there were few events in the training set is consistent with simulation studies that have shown increased bias and variability in parameter estimates when the number of events per variable (EPV) is low. A commonly

suggested guideline, and one supported by these simulation studies, is that with logistic regression EPV should be ten or greater [40]. Also using a simulation study approach, Vittinghoff and McColluch (2007) suggested that under some scenarios this ten EPV rule could be relaxed. However, they also noted that the EPV below which logistic regression had increasingly poor performance depended on data structure, and suggested that a higher EPV—such as EPV ≥ 20—might be appropriate in the prediction setting. In particular, they found binary predictors with low predictor prevalence necessitate a larger EPV [47]. Note that in the data used in this simulation almost all predictors were binary, and the predictor prevalence could be low.

With 149 predictors, the events per variable in our training set did not exceed 10 until $n_{\text{event}} = 1,837$ (EPV = 12.3), which corresponds to a training set size of $100,000$ and our largest event rate of 0.0184. Even under a more relaxed EPV rule of thumb, the EPV did not exceed five until $n_{\text{event}} = 919$, a number of events corresponding to $n_{\text{train}} = 100,000$ and $R = 0.0092$ or $n_{\text{train}} = 50,000$ and $R = 0.0046$. With the three event rates we considered in this simulation study, with a training set size of $10,000$ we always had an EPV < 2 and with a training set size of $5,000$ we always had an EPV < 1. We would expect random forest and ridge regression to have better performance when the EPV is low, since both algorithms perform variable selection. With ridge regression, this variable selection is implemented via the $L2$ penalty. While this $L2$ penalty does not result in variables being "kicked out" of the model as with the lasso and its $L1$ penalty, it shrinks the coefficients of some variables very close to zero. With random forest, we see an implementation of variable selection in the selection of the split rule, at each split, that results in the greatest classification improvement as measured by the Gini index [24].

Another sign of poor logistic regression performance with small n_{event} was that while the estimation procedure never failed to converge, with a small number of events we often encountered warnings that the fitted logistic regression model resulted in predicted probabilities of zero or one. While for the purposes of this simulation study we ignored these warnings, they are a sign that the model was over-fitting the data and that the predicted probabilities were unreliable.

6.2 Naive versus stratified sampling strategies for creating cross-validation folds

At the smallest training set size, $n_{\text{train}} = 5,000$, and two smallest event rates, a stratified sampling strategy for creating cross-validation folds sometimes resulted in slightly improved coverage of the bootstrap intervals and a lower MSE for AUC. When accounting for the Monte Carlo error present in this simulation study, there was little difference in absolute bias comparing a naive and stratified sampling strategy, even at the smallest training set size and event rates. However, the improved reliability provided by utilizing a stratified sampling strategy for creating cross-validation folds was neither consistent nor large in magnitude, and any advantage of a stratified sampling strategy disappeared once there were a moderate number of training set events.

The clearest advantage of a stratified sampling strategy for creating cross-validation folds when there are few training set events is that it ensures that all folds have at least one event. With few events, a naive sampling strategy often resulted in folds without any events. Calculating AUC requires at least one true event, and thus we could not calculate out-of-sample AUC for these folds with no events. While in this simulation study we worked around this issue by averaging over the out-of-sample AUC from only those folds with events to obtain a cross-validated AUC estimate, this solution is not desirable. It reduces the number of models averaged over to obtain the cross-validated estimate, and those models averaged over are trained on data that in aggregate has a slightly inflated event rate in comparison to the entire training set. In contrast, a stratified sampling strategy ensures that every fold has an event—unless there are more folds than events, in which case either there are unreasonably many folds or such a small number of events that the cross-validation sampling strategy becomes a secondary concern.

6.3 Ten-by-ten cross-validation and improvement in reliability

In contrast to using a stratified sampling strategy, ten-by-ten cross-validation offered a small but consistent improvement in reliability for small training set sizes and event rates. For all algorithms, the empirical MSE of the ten-by-ten cross-validated AUC for the iteration-

specific test set AUC was smaller than the MSE of the one-by-ten (one round of ten-fold cross-validation) cross-validated AUC (Figure 4.13). Additionally, for all algorithms the empirical coverage of the bootstrap intervals obtained using the out-of-sample predicted probabilities from ten rounds of ten-fold cross-validation was greater than the coverage of bootstrap intervals obtained using the predicted probabilities from only one round of ten-fold cross-validation (Figure 4.14). Ten-by-ten cross-validation also generally resulted in lower absolute empirical bias for AUC, although perhaps due to the larger relative Monte Carlo error for bias this improvement was not as consistent as that for MSE or coverage (Figure 4.12).

Improvement in AUC coverage, bias, and MSE was greatest for logistic regression and smallest for ridge regression, with improvement for the random forest algorithm intermediate. However, the improved reliability offered by averaging over the cross-validated estimates, or in the case of coverage bootstrapping with the out-of-sample predicted probabilities from multiple rounds of cross-validation, quickly decreased as the number of events in the training set increased. By $n_{\text{train}} = 10,000$ and our second largest event rate, $R = 0.0092$, the difference in reliability measures for ten-by-ten and one-by-ten was negligible. For some algorithms (particularly ridge regression) and measures of reliability, the difference became negligible even sooner.

These results from the simulation study suggest that multiple rounds of cross-validation to estimate performance measures and obtain hold-out-fold predicted probabilities can be advantageous in the rare event setting when the training set size is relatively small. However, there is little advantage to multiple rounds cross-validation when the training set size and number of events are larger. At larger training set sizes, doing many rounds of cross-validation is also more computationally costly.

6.4 Comparing confidence interval types

We observed little difference in empirical coverage with different types of bootstrap confidence intervals. There was little difference between Wald and percentile-type intervals. There was also little difference between re-sampling across the aggregated out-of-sample predicted probabilities obtained from the ten-fold cross validation versus re-sampling and

calculating the prediction metric within the folds, then averaging over the folds. This result suggests that there is little advantage to respecting the original cross-validation fold divisions in the bootstrapping procedure, as opposed to ignoring the folds and sampling the out-of-sample predicted probabilities in aggregate. We also find no compelling reason, based on this simulation study and the observed empirical coverage, to prefer either a Wald-type or percentile-type interval for AUC. However, we also note that Wald intervals may be undesirable because they assume that AUC estimates follow a normal distribution, which may not be the case. There are also other types of bootstrap intervals not considered in this simulation study, such as bootstrap intervals obtained using the studentised or non-studentised pivotal method, or bias-corrected percentile intervals [8].

Bootstrap intervals did provide better empirical coverage than influence curve based confidence intervals for AUC when the number of events in the training set was small. Not until a training set size of 50,000 did the influence curve based intervals provide similar empirical coverage to the bootstrap intervals. However, we would expect poor coverage of influence curve intervals when the training set size is small, since these intervals use an asymptotic estimate of the variance of the cross-validated AUC and thus are asymptotic 95% intervals. Fundamentally, the validity of these intervals arises from the Central Limit Theorem and the fact that the empirical AUC is an asymptotically linear and normal estimator for the true AUC [31]. In fact, in their paper proposing the influence curve based intervals for AUC, LeDell, Petersen, and van der Laan show, using a simulation study approach, that the coverage of these intervals may be below the nominal 95% level when the size of the training set is small. They suggest that with a small training set size bootstrapping may be a preferable for obtaining confidence intervals, especially because with a small training set size the computational expense of bootstrapping is not prohibitively high. The true advantage of the influence curve approach is that with a large training set size it provides an interval with good coverage without having to bootstrap [31]. Our results are consistent with this argument, since the influence curve intervals for AUC had equivalent coverage to the bootstrap intervals with large training set sizes and at large training set sizes bootstrapping was computationally expensive.

6.5 Impact of sampling a test set on the evaluation of empirical coverage

We were initially surprised by the drop in empirical coverage observed at the largest training set size, $n_{\text{train}} = 1$ million. For some metrics and algorithms, we even saw a drop in coverage at $n_{\text{train}} = 100,000$, in comparison to the empirical coverage at the next smallest training set size, $n_{\text{train}} = 50,000$. However, further investigation suggested that this decrease in empirical coverage was due to the additional variability introduced by sampling a test set of size 1 million. The iteration-specific test set AUC is not the true AUC parameter, rather it is itself an estimate of that parameter.

Using a fixed model trained on a test set of size 1 million, we found that the variance of test set AUC (from 1,000 independently sampled test sets) was approximately 0.0016. While this variability from test set sampling was present at all training set sizes, it only resulted in noticeably lower empirical coverage at large training set sizes because of the narrow width of the confidence intervals at these large training set sizes. When the width of the intervals is small, the relative variability due to sampling the test set is large, and thus becomes an important driver of the coverage of the iteration-specific test set prediction metric. However, when the width of the intervals is large relative to this variability, as is the case at smaller training set sizes, this variability has a small or negligible impact on empirical coverage.

Since the goal of this simulation study was to evaluate how, in a rare event setting, the reliability of AUC and other prediction metrics change as the size of the training set increases, the impact of this variability due to test set sampling on measures of reliability—especially confidence interval coverage—at larger training set sizes is an important limitation to our results. We used two approaches to account for this variability. Although we limited these approaches to AUC, they could also be applied to other performance metrics. While these approaches provided important insight into the impact of this test set variability, ultimately both have drawbacks.

The first way in which we accounted for the variability from sampling a test set was to look at the coverage of the average test set AUC, where we averaged over simulation replicates with the same training set size, sampling strategy, algorithm, and event rate

(Figure 4.3). While this averaging reduced variability due to sampling the test set, since each simulation replicate used a different training set each simulation replicate had a different fitted model. Thus, the coverage of this average test set AUC can be thought of as the coverage of the true AUC parameter for a given algorithm with given fixed hyperparameters, but not the coverage of the true AUC of a specific fitted model. While both true AUCs are interesting quantities, the quantity we were focused on in this simulation study was the true AUC for a specific fitted model.

The second way in which we accounted for the variability from sampling a test set was to look at the coverage of the iteration-specific test set AUC $\pm$ the approximate standard deviation of test set AUC from a single fixed model trained on a training set of size 1 million (Figure 4.11). We used $\text{SD}(\text{AUC}_{\text{test}}) = 0.0016$ for all algorithms, hyperparameters (which differed for a naive versus stratified sampling strategy), event rates, and training set sizes. However, the small simulation we conducted to evaluate the variability of test set AUC across test sets suggested that the variability may differ by algorithm (Table 4.4). The variability may also change with training set size, fixed hyperparameters, and event rate. Thus, using a single value for $\text{SD}(\text{AUC}_{\text{test}})$ may be problematic. More fundamentally, even with good, setting-dependent estimates of $\text{SD}(\text{AUC}_{\text{test}})$, adjusting for test set variability by examining the coverage of the iteration-specific test set AUC $\pm$ $\text{SD}(\text{AUC}_{\text{test}})$ is an imperfect solution. With this approach, we are looking at the overlap of two intervals, and for any one iteration the true AUC could be very close to the test set AUC or further from the test set AUC than $\text{SD}(\text{AUC}_{\text{test}})$.

Given the drawbacks of the two above approaches for adjusting for the variability due to test set sampling and the impact of this variability on the results of this simulation study, we plan to re-run the simulations with a small change in procedure. Instead of using a single test set to estimate the true model-specific prediction metric, m_{true}, for each replicate we will sample multiple test sets, calculate the metric from the fitted model in each test set, and then average over the test set metrics to obtain our estimate of the true out-of-sample prediction metric, $\hat{m}$. For example, with h test sets, our estimate of the true metric would

be

$$\hat{m} = \frac{1}{h}\sum_{i=1}^{h} m_{\text{test},i}.$$

Averaging over independently sampled test sets in this way will reduce the variability due to test set sampling and give us a better estimate of the true metric. We will then use this estimate $\hat{m}$ to obtain empirical coverage, empirical bias, and empirical MSE.

Before we re-run the simulations, we first need to establish how many test sets h we need to sample in each simulation replicate. In order to do so, we can use the widths of the 95% confidence intervals for the predictions metrics as a guide. We want the variability in $\hat{m}$ across iterations to be small relative to the width of the narrowest confidence intervals, but out of consideration of computation time we do not want the number of sampled test sets h to be unnecessarily large. A simulation-based approach for determining an appropriate h under this framework is proposed in Figure 6.1. In re-running the simulations, to cut down on computation we additionally will not do ten-by-ten cross-validation for the the two smallest training set sizes, $n_{\text{train}} = 5,000$ and $n_{\text{train}} = 10,000$, and we will only bootstrap across the aggregated holdout fold predicted probabilities. At the smallest training set sizes, the variability due to sampling the test set is small relative to the confidence interval width, so had little impact on our results. Thus, we would learn little by repeating the comparison of ten-by-ten and one-by-ten cross validation using the new estimate of true AUC. Ten-by-ten cross validation is also computationally costly, so overall it makes little sense to include it when re-running the simulations. Similarly, we saw little difference in the coverage of the types of bootstrap intervals, and doing only one type of bootstrapping will save computation time.

6.6 Reliance on n_{event}, $n_{\text{non-event}}$, and event rate

In overview, we found that the reliability of a prediction metric—as measured by empirical coverage of confidence intervals, empirical bias, empirical MSE, and variability of the cross-validated metric across replicates—could be driven by n_{event}, driven by $n_{\text{non-event}}$, or have some direct reliance on the event rate. Number of events in the training set drove the reliability of AUC and sensitivity. Number of non-events in the training set drove the

Simulation for determining the number of test sets h to sample

For each replicate

1. Sample a training set of size 1 million and 200 test sets of size 1 million
2. Fit the model on the training set
3. Using the fitted model, calculate each prediction metric m on each test set
4. Average over the first 10, first 25,..., all 200 test sets to obtain $\hat{m}_{10}$, $\hat{m}_{25}$,..., $\hat{m}_{200}$

For each algorithms/hyperparameter combination

5. Repeat 1-4 1,000 times
6. From the 1,000 replicates, calculate SD($\hat{m}_{10}$), SD($\hat{m}_{25}$), ..., SD($\hat{m}_{200}$)

Over all algorithms/hyperparameter combinations

7. For each metric and algorithm combination, using the previous simulation study results calculate w_m, the narrowest average CI width for metric m
8. Select the number of test sets h such that h is as small as possible but that, for all metrics m and algorithms/hyperparameter combinations, SD($\hat{m}_h$) $< g(w_m)$, where $g(w_m)$ is some function of w_m that serves as an appropriate threshold

Figure 6.1: Simulation procedure for deciding the number of test sets to sample when re-running the simulations. Note that the maximum variability across simulation replicates is a function of confidence interval width for the training set size and rate that results in the narrowest confidence intervals.

reliability of specificity. The reliability of accuracy, Brier score, PPV, NPV, F_1 score, and $F_{0.5}$ score all showed some direct reliance of event rate in addition to being driven by the size of the training set.

Of course, the number of events in a training set directly relies on the training set size and event rate: $n_{\text{event}} = n_{\text{train}} \times R$. When an event is rare, the number of events can be small even with relatively large data sets. However, if the reliability of a performance metric is driven by the number of events, with a large data set (e.g., millions of observations) even with a very rare event we will still have sufficiently many events for the estimated prediction metric to provide a reliable understanding of the performance of a predictive model. We do not have to worry about the prediction metric being unreliable simply because the event rate is small. While this distinction between poor performance of a prediction metric with respect to its reliability in rare event settings may be unimportant in settings where large data sets are not available, it is crucial in settings—such as the use of health-system EHR

data to build clinical prediction models—where large data sets are increasingly available and computationally feasible to work with.

6.7 AUC is reliable when number events is large

This last point is particularly salient with respect to AUC, which is commonly used to assess discriminative model performance. The results of this simulation study show that the reliability of AUC is driven by the number of events in the training set, not event rate. Moreover, even in the setting of a very rare event—the smallest event rate we considered was $R = 0.0046$, or approximately one event for every 217 non-events—these simulations show that AUC performs well provided that the number of events in the training set is sufficiently large. The largest training set size we considered was 1 million, which corresponds to 4,591 training set events for the smallest event rate. At this number of events, the cross-validated AUC was very nearly unbiased for the iteration-specific test set AUC (Figure 4.4), and the MSE was very near zero (Figure 5.3). The variability of the cross-validated AUC across simulation replicates was also small (Figure 4.8). The IQR of the cross-validated AUC was only 0.004 for ridge regression (under both a naive and stratified sampling strategy), and was slightly smaller for the logistic regression and random forest algorithms. While evaluating the empirical coverage of the confidence intervals was complicated by the variability introduced by sampling a test set, we hypothesize that when re-running the simulations—with the previously described modification meant to minimize that variability—we will see empirical coverage at the nominal 95% level.

We did find that AUC was less reliable when the number of events in the training set was very small. In particular, there was notably high variability in cross-validated AUC across sampled training sets (Figure 4.8), and the distance between the cross-validated AUC and the test set AUC for any one test set/training set pair could be large (Figure 4.5). For example, with ridge regression and the the second smallest number of events, $n_{\text{event}} = 46$, the IQR of the cross-validated AUC estimates was approximately 0.05 for all sampling strategies and event rates, while the range was about 0.25. While the poor performance of AUC with so few events may be expected, we feel that it is still important to highlight here. However, even at our very smallest event rate, $R = 0.0046$, we did not need a test set of size

1 million for the cross-validated AUC to show good reliability. At this smallest event rate cross-validated AUC was nearly unbiased for the iteration test-set AUC (bias < 0.001), had low MSE (< 0.0002), and had low variability (IQR for ≤ 0.015) across iterations even when there were only 460 events in the training set ($n_{\text{train}} = 100,000$). While the acceptable amounts of bias and variability are clearly context dependent, overall we conclude that AUC has good performance in the rare event setting when that the total number of events is moderately large.

Chapter 7

CONCLUSION

This simulation study shows that the reliability of AUC is driven by the number of events in the training set, not the event rate. In contrast, some other commonly used metrics of model performance, especially accuracy, have reliability that is in part driven by the event rate. Even in the rare event setting, we found AUC to be reliable as long as the number of events in the training set was sufficiently large. However, while AUC is both a useful measure of the performance of a predictive model and is reliable in the rare event setting, we also emphasize that looking at multiple metrics of model performance can give a more complete picture of a model's strengths and weaknesses. Different performance metrics provide different information, and ultimately which metric is of greatest interest depends on how a predictive model will be used.

BIBLIOGRAPHY

[1] N. M. Adams and D. J. Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern recognition*, 32(7):1139–1147, 1999. Place: Oxford Publisher: Elsevier Ltd.

[2] Samrachana Adhikari, Sharon-Lise Normand, Jordan Bloom, David Shahian, and Sherri Rose. Revisiting performance metrics for prediction with rare outcomes. *Statistical Methods in Medical Research*, 30(10):2352–2366, October 2021. Publisher: SAGE Publications Ltd STM.

[3] Peter C Austin and Ewout W Steyerberg. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*, 26(2):796–808, April 2017.

[4] Bradley E. Belsher, Derek J. Smolenski, Larry D. Pruitt, Nigel E. Bush, Erin H. Beech, Don E. Workman, Rebecca L. Morgan, Daniel P. Evatt, Jennifer Tucker, and Nancy A. Skopp. Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*, 76(6):642–651, June 2019.

[5] Simon Bernard, Laurent Heutte, and Sébastien Adam. Influence of Hyperparameters on Random Forest Accuracy. In Jón Atli Benediktsson, Josef Kittler, and Fabio Roli, editors, *Multiple Classifier Systems*, Lecture Notes in Computer Science, pages 171–180, Berlin, Heidelberg, 2009. Springer.

[6] Ehsan Bokhari. Clinical (In)Efficiency in the Prediction of Dangerous Behavior. *Journal of educational and behavioral statistics*, pages 107699862211447–, 2023. Place: Los Angeles, CA Publisher: SAGE Publications.

[7] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. Place: Oxford Publisher: Elsevier Ltd.

[8] J. Carpenter and J. Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19(9):1141–1164, May 2000.

[9] Jonathan H. Chen and Steven M. Asch. Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *The New England journal of medicine*, 376(26):2507–2509, June 2017.

[10] R. Yates Coley, Eric Johnson, Gregory E. Simon, Maricela Cruz, and Susan M. Shortreed. Racial/Ethnic Disparities in the Performance of Prediction Models for Death by Suicide After Mental Health Visits. *JAMA Psychiatry*, 78(7):726–734, July 2021.

[11] Nancy R. Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, February 2007.

[12] Corinna Cortes and Mehryar Mohri. AUC optimization vs. Error rate minimization: 17th Annual Conference on Neural Information Processing Systems, NIPS 2003. *Advances in Neural Information Processing Systems 16 - Proceedings of the 2003 Conference, NIPS 2003*, 2004. Publisher: Neural information processing systems foundation.

[13] Folashade Daniel, Hong Ooi, Rich Calaway, Microsoft, and Steve Weston. foreach: Provides Foreach Looping Construct, February 2022.

[14] George A. Diamond. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology*, 45(1):85–89, January 1992.

[15] Lori E. Dodd and Margaret S. Pepe. Partial AUC Estimation and Regression. *Biometrics*, 59(3):614–623, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1541-0420.00071.

[16] Seena Fazel, Matthias Burghart, Thomas Fanshawe, Sharon Danielle Gil, John Monahan, and Rongqin Yu. The predictive performance of criminal risk assessment tools used at sentencing: Systematic review of validation studies. *Journal of Criminal Justice*, 81:101902, July 2022.

[17] Jerome Friedman, Robert Tibshirani, and Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[18] David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, October 2009.

[19] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982. Publisher: Radiological Society of North America.

[20] J A Hanley and B J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, September 1983. Publisher: Radiological Society of North America.

[21] Trevor Hastie. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, second edition, corrected 7th printing. edition, 2009.

[22] L. Maaike Helmus and Kelly M. Babchishin. Primer on Risk Assessment and the Statistics Used to Evaluate Its Accuracy. *Criminal Justice and Behavior*, 44(1):8–25, January 2017. Publisher: SAGE Publications Inc.

[23] Jin Huang and C.X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, March 2005. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer, New York, NY, 2013.

[25] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear Model Selection and Regularization. In Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors, *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, pages 203–264. Springer, New York, NY, 2013.

[26] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Tree-Based Methods. In Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors, *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, pages 303–335. Springer, New York, NY, 2013.

[27] WILLIAM A. KNAUS, ELIZABETH A. DRAPER, DOUGLAS P. WAGNER, and JACK E. ZIMMERMAN. APACHE II: A severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985. Place: Hagerstown, MD Publisher: Williams & Wilkins.

[28] Kenneth D. Kochanek, Jiaquan Xu, and Elizabeth Arias. Mortality in the United States, 2019. *NCHS data brief*, (395):1–8, 2020. Place: United States.

[29] Elizabeth Koehler, Elizabeth Brown, and Sebastien J.-P. A. Haneuse. On the assessment of monte carlo error in simulation-based statistical analyses. *The American statistician*, 63(2):155–162, 2009.

[30] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173, April 2009.

[31] Erin LeDell, Maya Petersen, and Mark van der Laan. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electronic journal of statistics*, 9(1):1583–1607, 2015.

[32] Jake Lever, Martin Krzywinski, and Naomi Altman. Classification evaluation. *Nature Methods*, 13(8):603–604, August 2016. Number: 8 Publisher: Nature Publishing Group.

[33] Charles X. Ling, Jin Huang, and Harry Zhang. AUC: a better measure than accuracy in comparing learning algorithms. In *Proceedings of the 16th Canadian society for computational studies of intelligence conference on Advances in artificial intelligence*, AI'03, pages 329–341, Berlin, Heidelberg, June 2003. Springer-Verlag.

[34] Barbara J. McNeil and James A. Hanley. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making*, 4(2):137–150, June 1984. Publisher: SAGE Publications Inc STM.

[35] Charles E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298, October 1978.

[36] Karel G. M. Moons, Joris A. H. de Groot, Walter Bouwmeester, Yvonne Vergouwe, Susan Mallett, Douglas G. Altman, Johannes B. Reitsma, and Gary S. Collins. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. *PLOS Medicine*, 11(10):e1001744, October 2014. Publisher: Public Library of Science.

[37] Karel G. M. Moons and Frank E. Harrell. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology*, 10(6):670–672, June 2003.

[38] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*, 375(13):1216–1219, September 2016.

[39] Michael J. Patton and Vincent X. Liu. Predictive Modeling Using Artificial Intelligence and Machine Learning Algorithms on Electronic Health Record Data. *Critical Care Clinics*, page S074907042300009X, April 2023.

[40] Peter Peduzzi, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379, December 1996.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[42] Saharon Rosset. Model selection via the AUC. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, page 89, New York, NY, USA, July 2004. Association for Computing Machinery.

[43] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE*, 10(3):e0118432, March 2015.

[44] Gregory E. Simon, Eric Johnson, Jean M. Lawrence, Rebecca C. Rossom, Brian Ahmedani, Frances L. Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B. Penfold, and Susan M. Shortreed. Predicting Suicide Attempts and Suicide Deaths Following Outpatient Visits Using Electronic Health Records. *American Journal of Psychiatry*, 175(10):951–960, October 2018. Publisher: American Psychiatric Publishing.

[45] Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128–138, January 2010.

[46] John A. Swets. Measuring the Accuracy of Diagnostic Systems. *Science*, 240(4857):1285–1293, 1988. Publisher: American Association for the Advancement of Science.

[47] Eric Vittinghoff and Charles E. McCulloch. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology*, 165(6):710–718, March 2007.

[48] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1171–1180, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee.

Appendix A

SIMULATION PROCEDURE DIAGRAMS

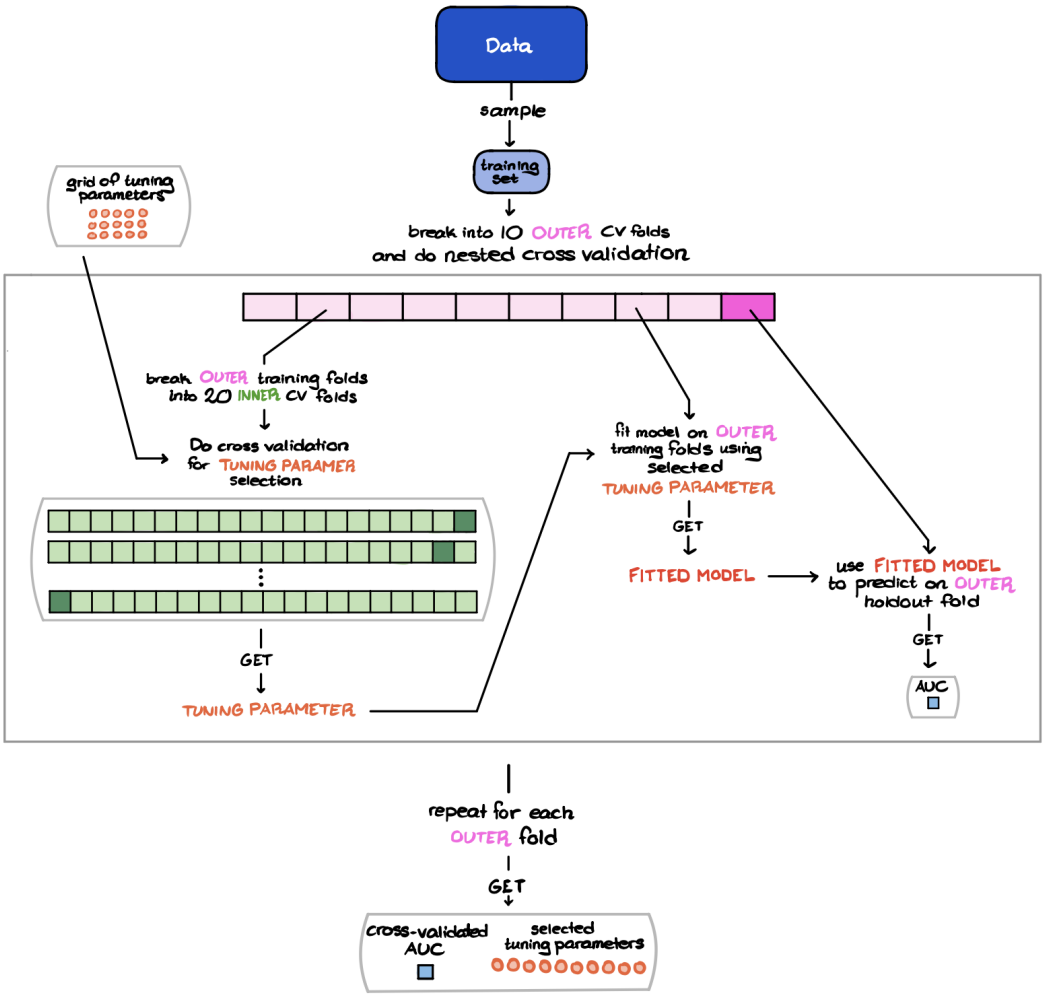

Figure A.1: Nested cross-validation

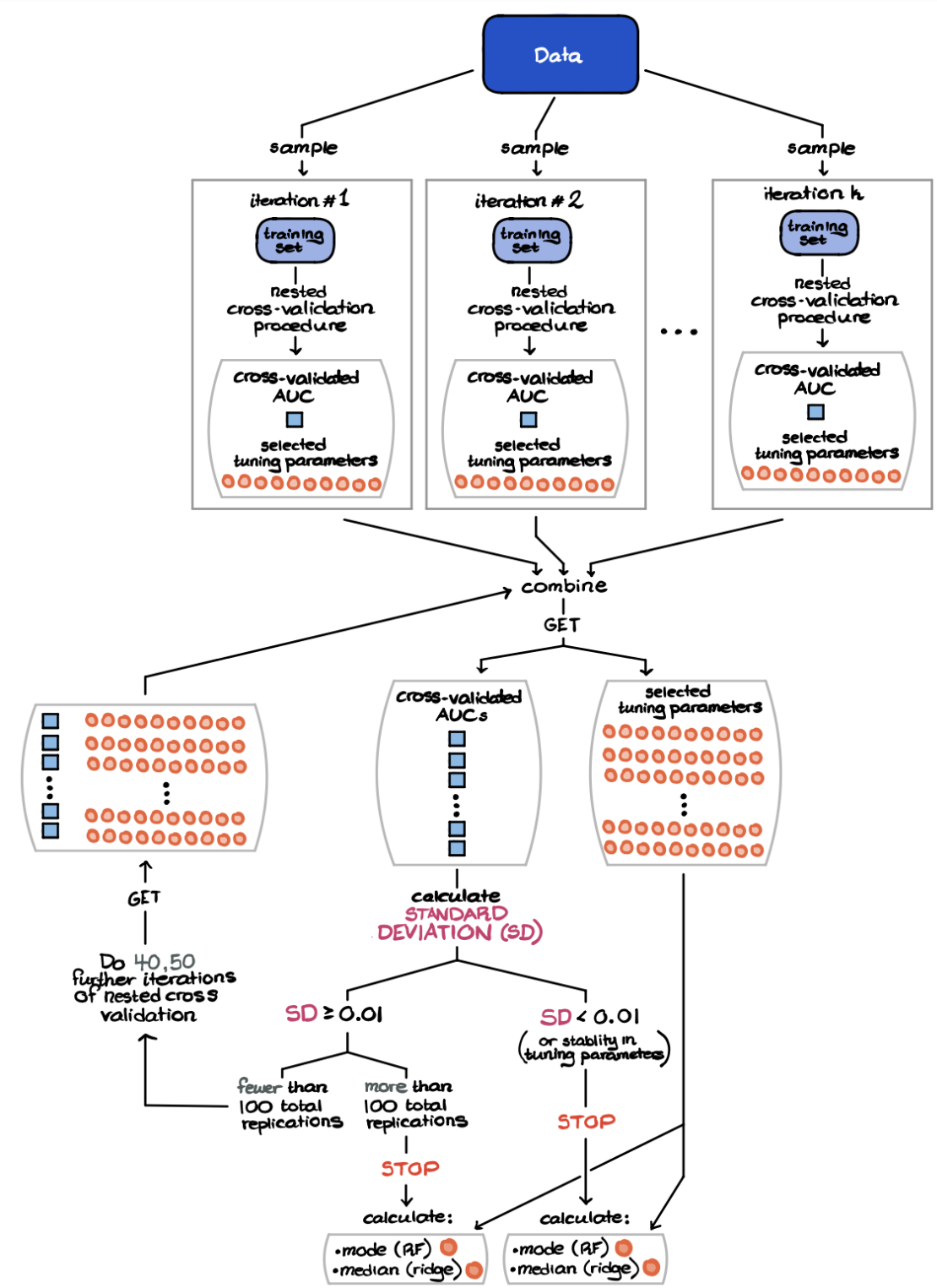

Figure A.2: Tuning parameter selection procedure

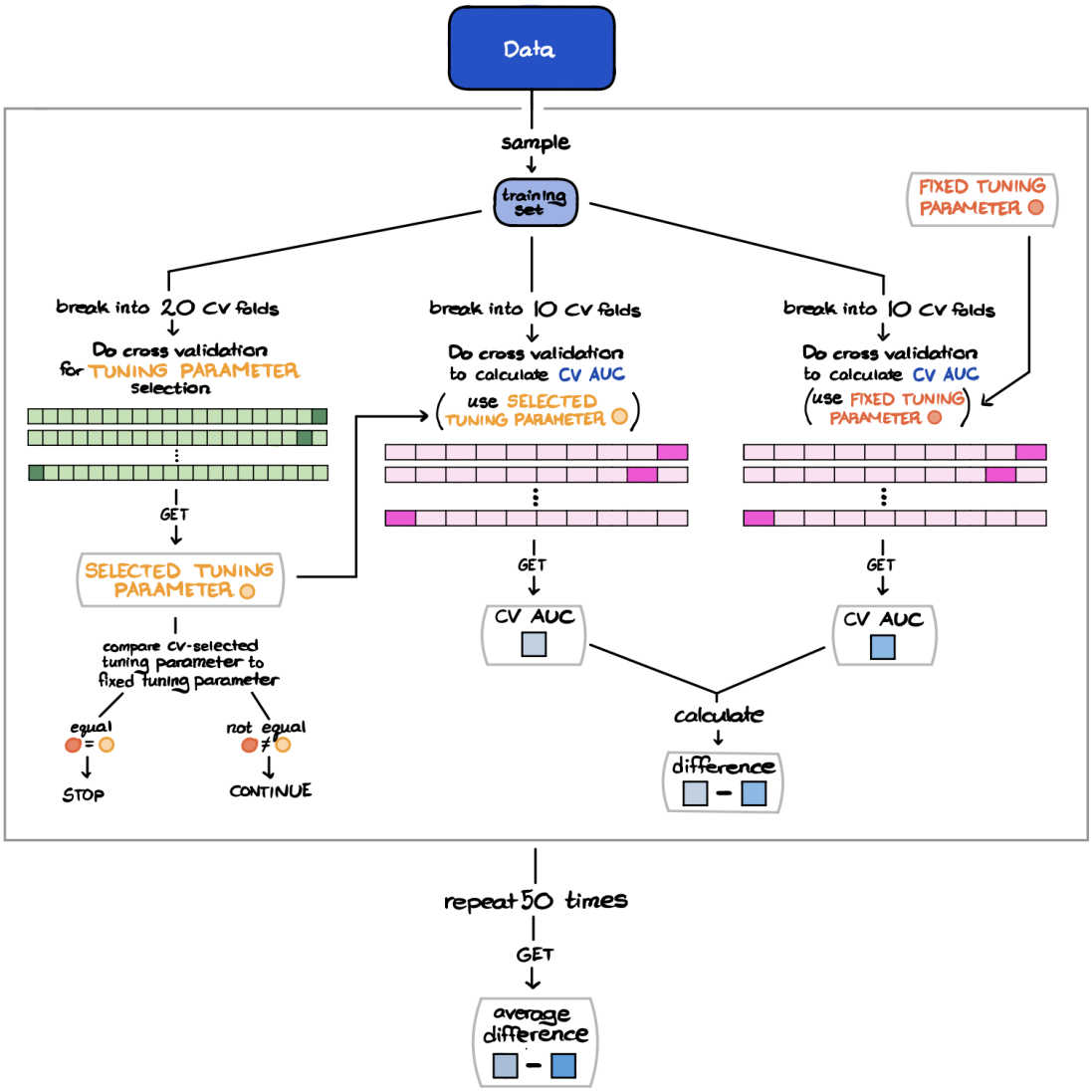

Figure A.3: Pilot simulation

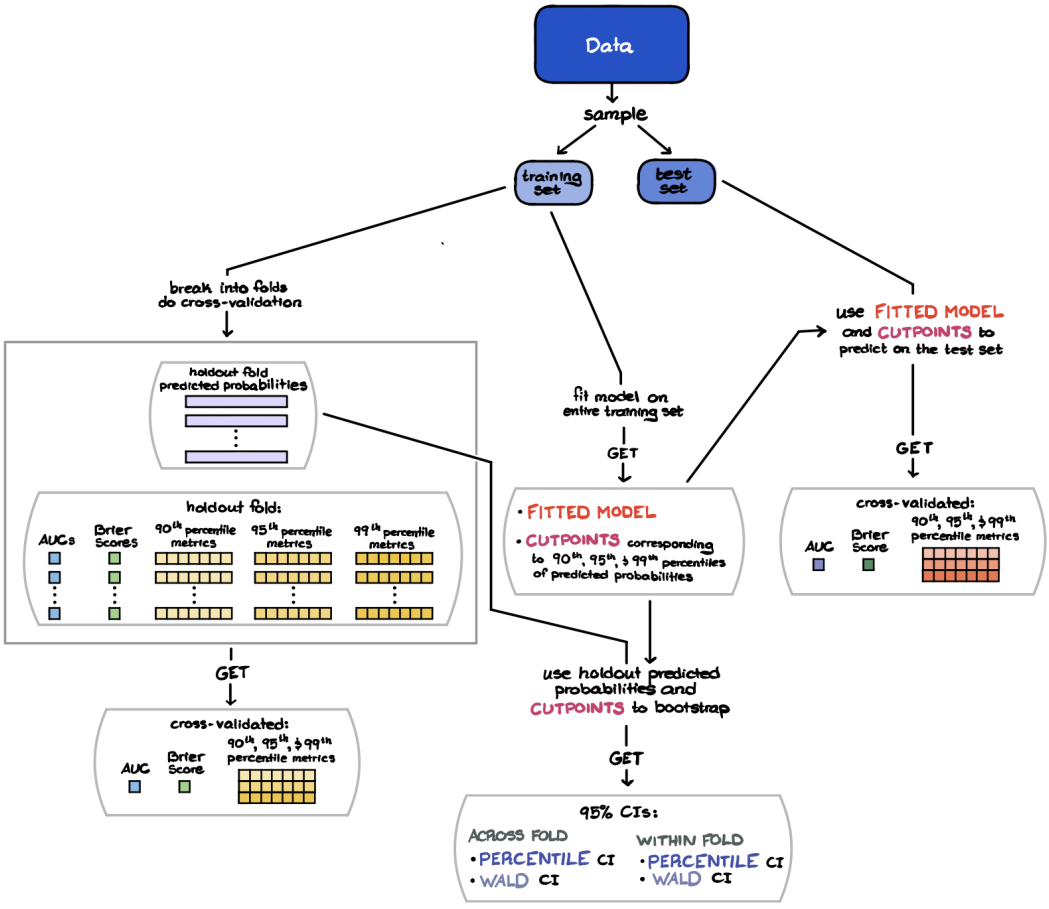

Figure A.4: Main simulation, overview of procedure

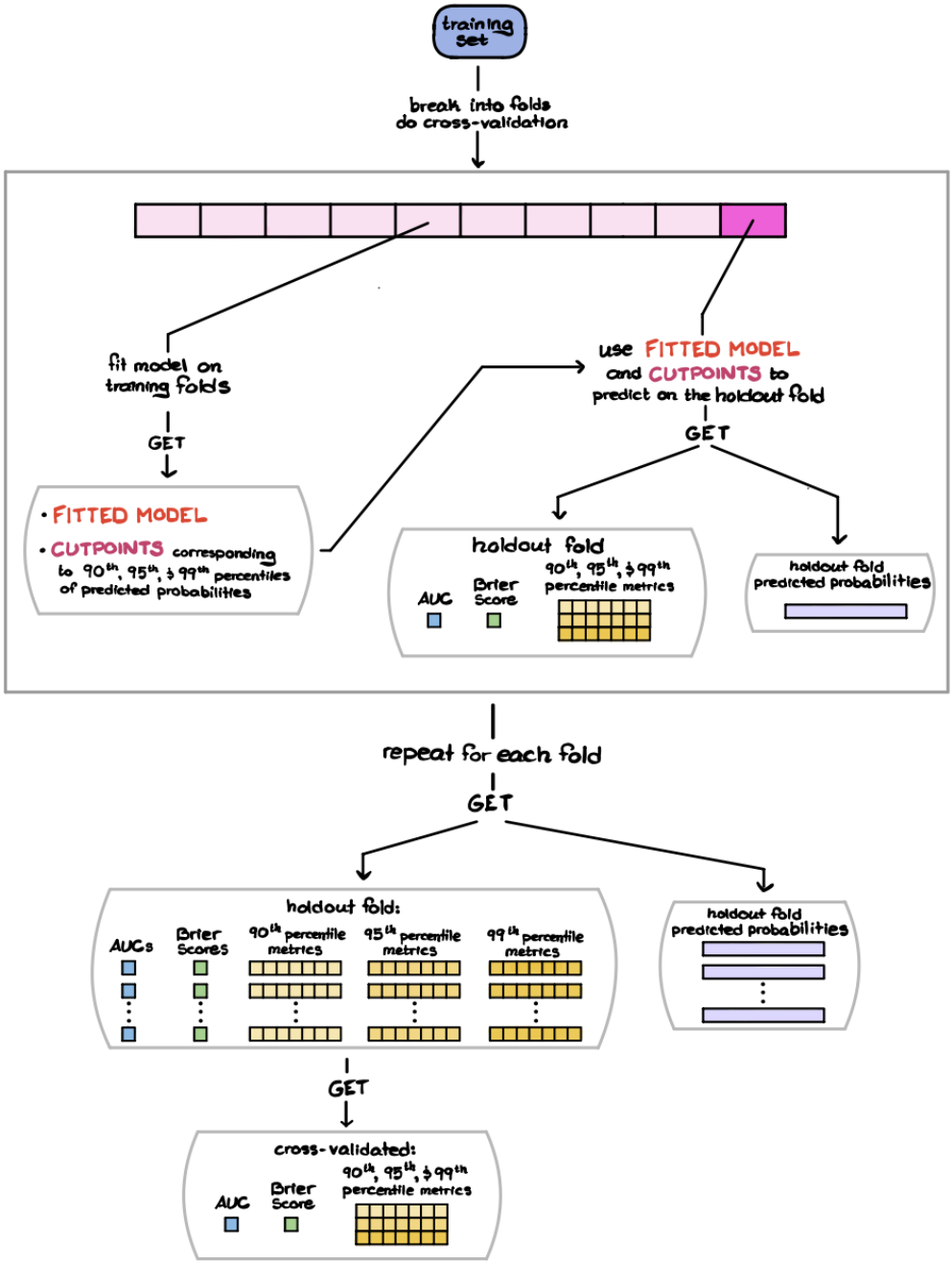

Figure A.5: Main simulation, cross-validation

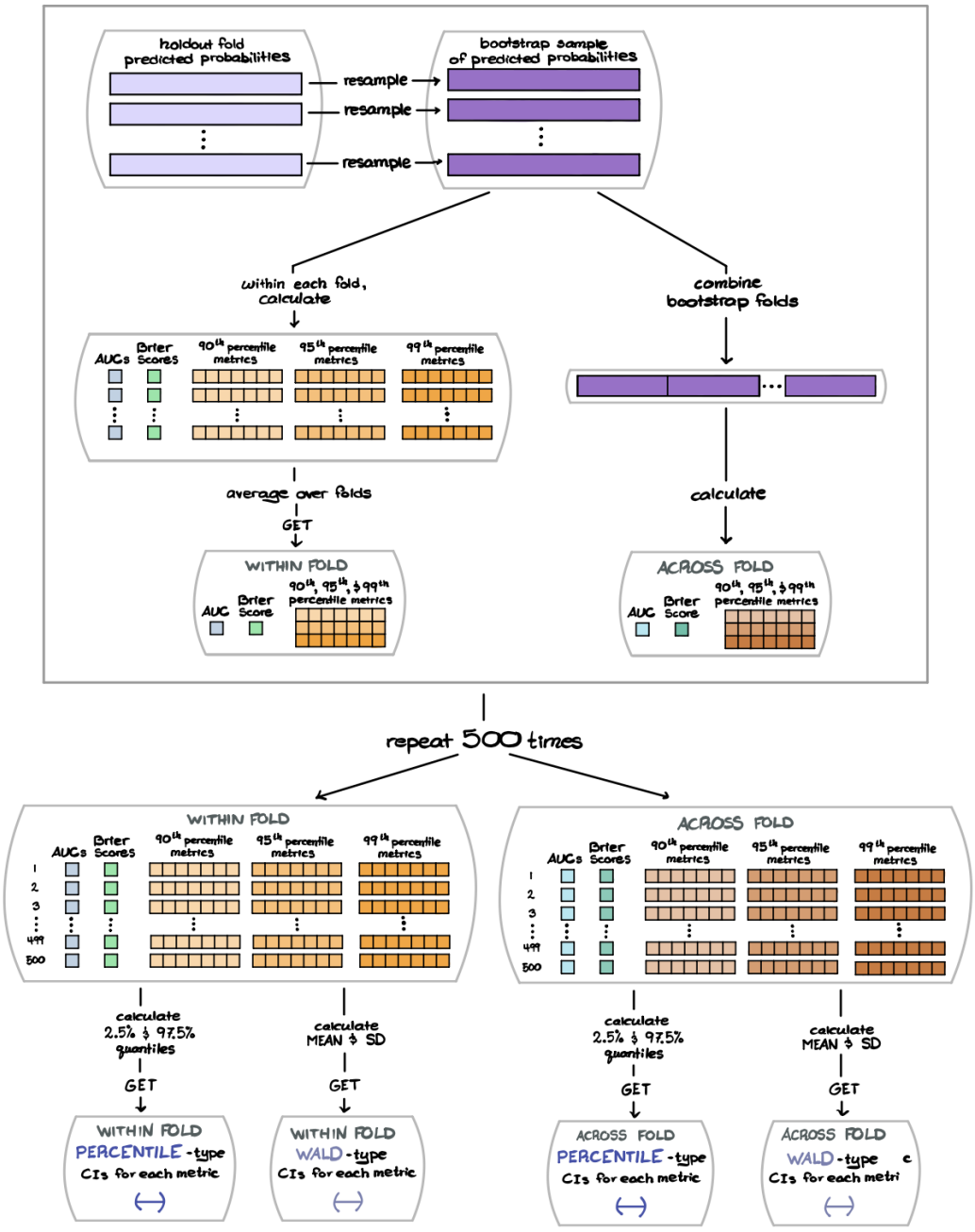

Figure A.6: Main simulation, bootstrapping procedure

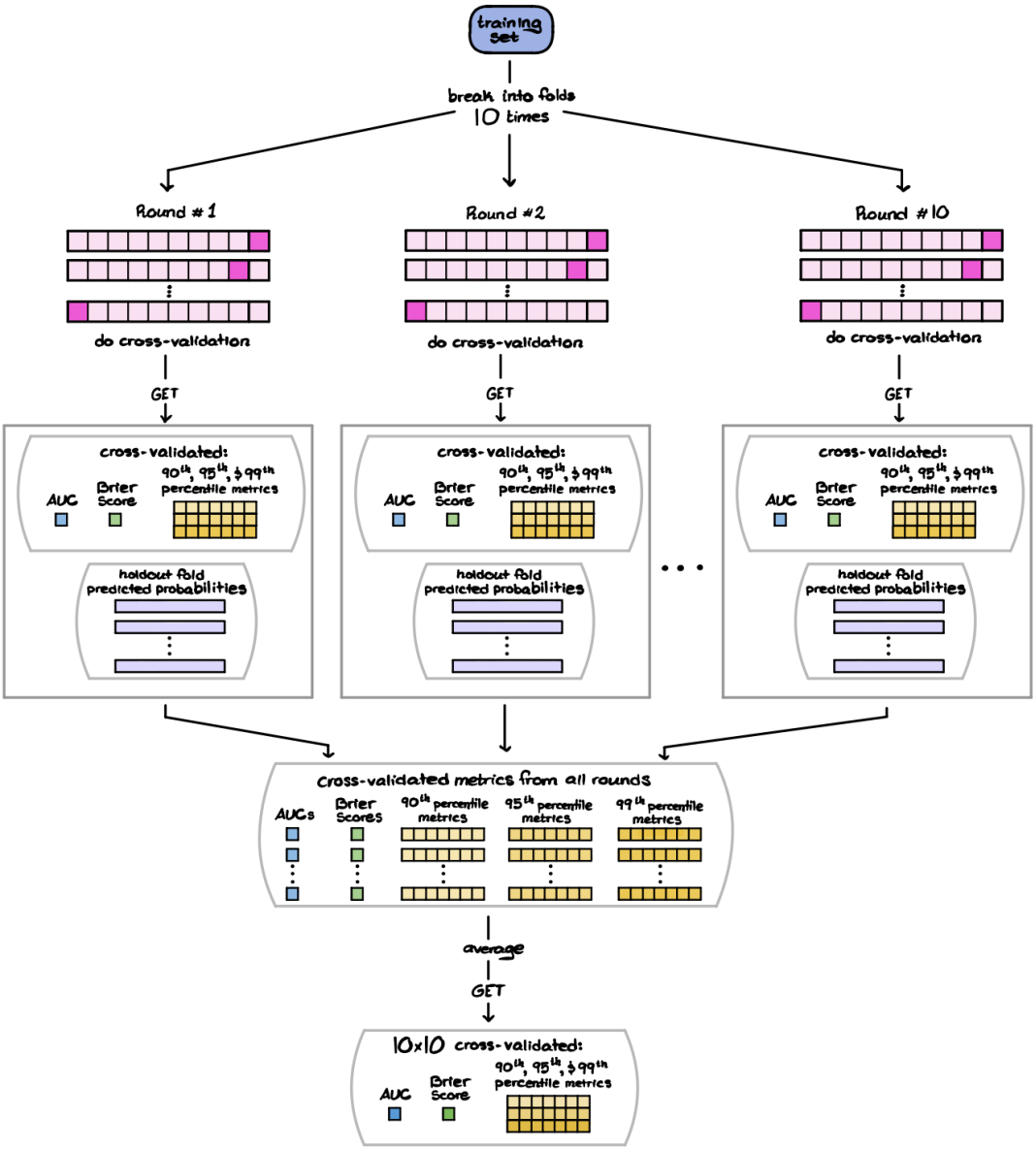

Figure A.7: Ten-by-ten cross-validation

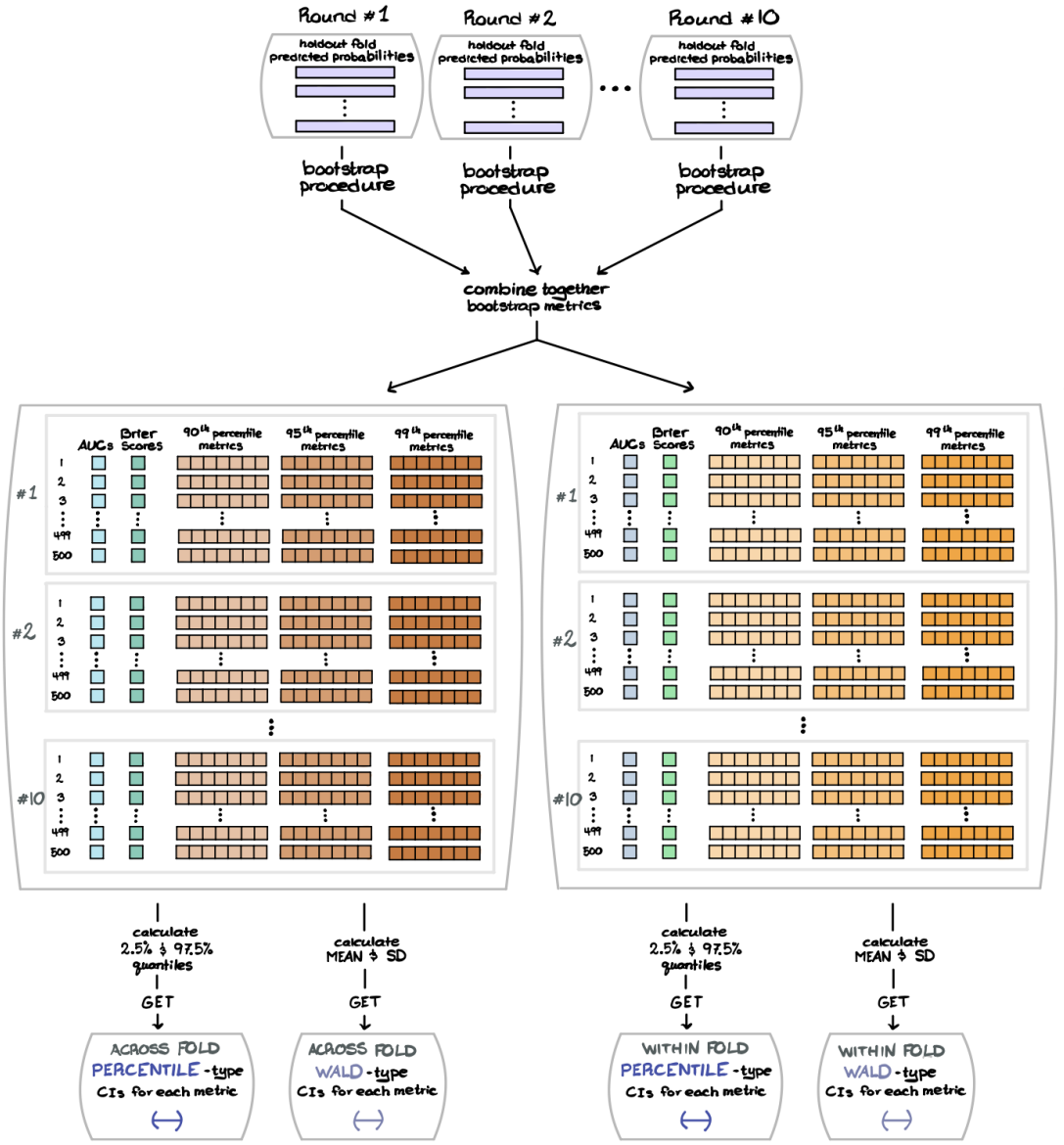

Figure A.8: Ten-by-ten bootstrapping procedure

Appendix B

CONFIDENCE INTERVAL COVERAGE

B.1 Area under the ROC curve (AUC)

			Coverage				
Training set size	Rate	Sampling strategy	CV	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	0.752	0.904	0.868	0.912	0.870
5,000	0.0046	stratified	0.770	0.880	0.848	0.892	0.854
5,000	0.0092	naive	0.816	0.918	0.884	0.920	0.894
5,000	0.0092	stratified	0.890	0.946	0.938	0.932	0.938
5,000	0.0184	naive	0.902	0.950	0.920	0.944	0.926
5,000	0.0184	stratified	0.924	0.940	0.932	0.938	0.942
10,000	0.0046	naive	0.816	0.916	0.892	0.920	0.900
10,000	0.0046	stratified	0.892	0.934	0.926	0.932	0.936
10,000	0.0092	naive	0.878	0.926	0.920	0.924	0.916
10,000	0.0092	stratified	0.902	0.924	0.922	0.918	0.926
10,000	0.0184	naive	0.928	0.942	0.946	0.934	0.934
10,000	0.0184	stratified	0.934	0.944	0.940	0.938	0.942
50,000	0.0046	naive	0.928	0.930	0.930	0.930	0.930
50,000	0.0046	stratified	0.934	0.948	0.938	0.946	0.936
50,000	0.0092	naive	0.928	0.936	0.932	0.934	0.928
50,000	0.0092	stratified	0.922	0.922	0.922	0.916	0.922
50,000	0.0184	naive	0.930	0.938	0.928	0.934	0.932
50,000	0.0184	stratified	0.942	0.942	0.940	0.938	0.940
100,000	0.0046	naive	0.940	0.950	0.948	0.952	0.954
100,000	0.0046	stratified	0.934	0.934	0.934	0.940	0.934
100,000	0.0092	naive	0.940	0.944	0.936	0.936	0.936
100,000	0.0092	stratified	0.912	0.912	0.910	0.914	0.912
100,000	0.0184	naive	0.922	0.916	0.914	0.916	0.916
100,000	0.0184	stratified	0.906	0.908	0.912	0.902	0.902
1,000,000	0.0046	naive	0.836	0.842	0.830	0.834	0.832
1,000,000	0.0046	stratified	0.840	0.834	0.840	0.830	0.838
1,000,000	0.0092	naive	0.846	0.840	0.850	0.838	0.846
1,000,000	0.0092	stratified	0.830	0.830	0.834	0.826	0.832

Table B.1: Coverage of 95% confidence intervals for AUC, random forest.

			Coverage				
Training set size	Rate	Sampling strategy	CV	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	0.783	0.848	0.884	0.829	0.893
5,000	0.0046	stratified	0.798	0.912	0.878	0.907	0.876
5,000	0.0092	naive	0.837	0.912	0.905	0.910	0.918
5,000	0.0092	stratified	0.898	0.937	0.935	0.933	0.939
5,000	0.0184	naive	0.902	0.938	0.933	0.939	0.936
5,000	0.0184	stratified	0.922	0.943	0.944	0.943	0.942
10,000	0.0046	naive	0.820	0.905	0.889	0.902	0.890
10,000	0.0046	stratified	0.891	0.934	0.926	0.937	0.924
10,000	0.0092	naive	0.925	0.955	0.951	0.948	0.953
10,000	0.0092	stratified	0.906	0.937	0.940	0.935	0.935
10,000	0.0184	naive	0.920	0.933	0.937	0.927	0.932
10,000	0.0184	stratified	0.935	0.942	0.943	0.934	0.944
50,000	0.0046	naive	0.909	0.928	0.927	0.920	0.925
50,000	0.0046	stratified	0.939	0.952	0.947	0.949	0.945
50,000	0.0092	naive	0.925	0.930	0.929	0.925	0.925
50,000	0.0092	stratified	0.940	0.945	0.947	0.938	0.938
50,000	0.0184	naive	0.943	0.944	0.944	0.938	0.942
50,000	0.0184	stratified	0.943	0.939	0.943	0.935	0.943
100,000	0.0046	naive	0.934	0.936	0.937	0.930	0.940
100,000	0.0046	stratified	0.909	0.916	0.914	0.914	0.909
100,000	0.0092	naive	0.934	0.930	0.935	0.928	0.935
100,000	0.0092	stratified	0.930	0.933	0.931	0.932	0.929
100,000	0.0184	naive	0.932	0.934	0.933	0.928	0.931
100,000	0.0184	stratified	0.931	0.928	0.933	0.927	0.924
1,000,000	0.0046	naive	0.824	0.820	0.819	0.819	0.817
1,000,000	0.0046	stratified	0.822	0.820	0.818	0.815	0.812
1,000,000	0.0092	naive	0.857	0.852	0.852	0.844	0.844
1,000,000	0.0092	stratified	0.853	0.851	0.854	0.842	0.851

Table B.2: Coverage of 95% confidence intervals for AUC, ridge regression.

			Coverage				
Training set size	Rate	Sampling strategy	CV	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	0.991	0.874	0.844	0.874	0.843
5,000	0.0046	stratified	0.991	0.865	0.821	0.861	0.821
5,000	0.0092	naive	0.833	0.805	0.865	0.806	0.863
5,000	0.0092	stratified	0.828	0.828	0.831	0.835	0.832
5,000	0.0184	naive	0.777	0.776	0.816	0.764	0.818
5,000	0.0184	stratified	0.763	0.763	0.784	0.749	0.781
10,000	0.0046	naive	0.734	0.805	0.814	0.804	0.809
10,000	0.0046	stratified	0.804	0.836	0.849	0.842	0.848
10,000	0.0092	naive	0.779	0.784	0.826	0.772	0.820
10,000	0.0092	stratified	0.816	0.825	0.840	0.821	0.833
10,000	0.0184	naive	0.819	0.818	0.836	0.816	0.835
10,000	0.0184	stratified	0.813	0.820	0.824	0.808	0.808
50,000	0.0046	naive	0.823	0.829	0.844	0.819	0.832
50,000	0.0046	stratified	0.834	0.845	0.847	0.834	0.835
50,000	0.0092	naive	0.874	0.871	0.882	0.862	0.871
50,000	0.0092	stratified	0.880	0.882	0.886	0.871	0.880
50,000	0.0184	naive	0.912	0.908	0.914	0.908	0.910
50,000	0.0184	stratified	0.920	0.921	0.916	0.910	0.914
100,000	0.0046	naive	0.878	0.873	0.888	0.869	0.880
100,000	0.0046	stratified	0.874	0.874	0.878	0.864	0.872
100,000	0.0092	naive	0.914	0.909	0.912	0.908	0.909
100,000	0.0092	stratified	0.897	0.899	0.900	0.898	0.891
100,000	0.0184	naive	0.940	0.935	0.940	0.927	0.939
100,000	0.0184	stratified	0.941	0.938	0.942	0.942	0.935
1,000,000	0.0046	naive	0.841	0.843	0.836	0.832	0.831
1,000,000	0.0046	stratified	0.822	0.818	0.817	0.817	0.816
1,000,000	0.0092	naive	0.819	0.816	0.819	0.811	0.811
1,000,000	0.0092	stratified	0.841	0.839	0.842	0.835	0.837

Table B.3: Coverage of 95% confidence intervals for AUC, logistic regression.

B.2 Brier Score

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	Brier	0.996	0.996	0.994	0.996
5,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
5,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
5,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
5,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
5,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
10,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
10,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
10,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
10,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
10,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
10,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000

Table B.4: Coverage of 95% confidence intervals for Brier score, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	Brier	1	0.999	0.999	0.999
5,000	0.0046	stratified	Brier	1	1.000	1.000	1.000
5,000	0.0092	naive	Brier	1	1.000	1.000	1.000
5,000	0.0092	stratified	Brier	1	1.000	1.000	1.000
5,000	0.0184	naive	Brier	1	1.000	1.000	1.000
5,000	0.0184	stratified	Brier	1	1.000	1.000	1.000
10,000	0.0046	naive	Brier	1	1.000	1.000	1.000
10,000	0.0046	stratified	Brier	1	1.000	1.000	1.000
10,000	0.0092	naive	Brier	1	1.000	1.000	1.000
10,000	0.0092	stratified	Brier	1	1.000	1.000	1.000
10,000	0.0184	naive	Brier	1	1.000	1.000	1.000
10,000	0.0184	stratified	Brier	1	1.000	1.000	1.000
50,000	0.0046	naive	Brier	1	1.000	1.000	1.000
50,000	0.0046	stratified	Brier	1	1.000	1.000	1.000
50,000	0.0092	naive	Brier	1	1.000	1.000	1.000
50,000	0.0092	stratified	Brier	1	1.000	1.000	1.000
50,000	0.0184	naive	Brier	1	1.000	1.000	1.000
50,000	0.0184	stratified	Brier	1	1.000	1.000	1.000
100,000	0.0046	naive	Brier	1	1.000	1.000	1.000
100,000	0.0046	stratified	Brier	1	1.000	1.000	1.000
100,000	0.0092	naive	Brier	1	1.000	1.000	1.000
100,000	0.0092	stratified	Brier	1	1.000	1.000	1.000
100,000	0.0184	naive	Brier	1	1.000	1.000	1.000
100,000	0.0184	stratified	Brier	1	1.000	1.000	1.000
1,000,000	0.0046	naive	Brier	1	1.000	1.000	1.000
1,000,000	0.0046	stratified	Brier	1	1.000	1.000	1.000
1,000,000	0.0092	naive	Brier	1	1.000	1.000	1.000
1,000,000	0.0092	stratified	Brier	1	1.000	1.000	1.000

Table B.5: Coverage of 95% confidence intervals for Brier score, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	Brier	0.403	0.396	0.392	0.391
5,000	0.0046	stratified	Brier	0.432	0.430	0.422	0.423
5,000	0.0092	naive	Brier	0.816	0.815	0.798	0.803
5,000	0.0092	stratified	Brier	0.837	0.842	0.828	0.831
5,000	0.0184	naive	Brier	0.971	0.970	0.969	0.966
5,000	0.0184	stratified	Brier	0.951	0.948	0.946	0.944
10,000	0.0046	naive	Brier	0.969	0.970	0.963	0.965
10,000	0.0046	stratified	Brier	0.959	0.953	0.953	0.952
10,000	0.0092	naive	Brier	0.993	0.993	0.993	0.993
10,000	0.0092	stratified	Brier	0.989	0.989	0.989	0.989
10,000	0.0184	naive	Brier	0.999	0.999	0.999	0.999
10,000	0.0184	stratified	Brier	0.999	0.999	0.997	0.998
50,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000

Table B.6: Coverage of 95% confidence intervals for Brier score, logistic regression.

B.3 Accuracy

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 90%	0.890	0.907	0.892	0.902
5,000	0.0046	stratified	accuracy 90%	0.926	0.920	0.922	0.918
5,000	0.0092	naive	accuracy 90%	0.938	0.939	0.932	0.934
5,000	0.0092	stratified	accuracy 90%	0.934	0.926	0.922	0.920
5,000	0.0184	naive	accuracy 90%	0.920	0.920	0.918	0.920
5,000	0.0184	stratified	accuracy 90%	0.952	0.946	0.946	0.940
10,000	0.0046	naive	accuracy 90%	0.876	0.884	0.876	0.874
10,000	0.0046	stratified	accuracy 90%	0.820	0.826	0.818	0.824
10,000	0.0092	naive	accuracy 90%	0.930	0.922	0.926	0.914
10,000	0.0092	stratified	accuracy 90%	0.916	0.914	0.906	0.910
10,000	0.0184	naive	accuracy 90%	0.946	0.942	0.942	0.936
10,000	0.0184	stratified	accuracy 90%	0.942	0.938	0.940	0.930
50,000	0.0046	naive	accuracy 90%	0.834	0.838	0.832	0.834
50,000	0.0046	stratified	accuracy 90%	0.844	0.836	0.842	0.834
50,000	0.0092	naive	accuracy 90%	0.848	0.844	0.840	0.846
50,000	0.0092	stratified	accuracy 90%	0.870	0.866	0.862	0.862
50,000	0.0184	naive	accuracy 90%	0.876	0.892	0.866	0.884
50,000	0.0184	stratified	accuracy 90%	0.880	0.884	0.878	0.884
100,000	0.0046	naive	accuracy 90%	0.772	0.780	0.770	0.776
100,000	0.0046	stratified	accuracy 90%	0.776	0.774	0.768	0.772
100,000	0.0092	naive	accuracy 90%	0.814	0.810	0.814	0.814
100,000	0.0092	stratified	accuracy 90%	0.816	0.804	0.802	0.810
100,000	0.0184	naive	accuracy 90%	0.804	0.804	0.804	0.804
100,000	0.0184	stratified	accuracy 90%	0.824	0.824	0.826	0.824
1,000,000	0.0046	naive	accuracy 90%	0.376	0.378	0.368	0.374
1,000,000	0.0046	stratified	accuracy 90%	0.340	0.340	0.332	0.334
1,000,000	0.0092	naive	accuracy 90%	0.306	0.312	0.306	0.312
1,000,000	0.0092	stratified	accuracy 90%	0.310	0.312	0.304	0.302

Table B.7: Coverage of 95% confidence intervals for accuracy at the 90th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 95%	0.912	0.918	0.918	0.934
5,000	0.0046	stratified	accuracy 95%	0.924	0.922	0.918	0.916
5,000	0.0092	naive	accuracy 95%	0.948	0.947	0.940	0.932
5,000	0.0092	stratified	accuracy 95%	0.934	0.942	0.926	0.938
5,000	0.0184	naive	accuracy 95%	0.942	0.946	0.938	0.946
5,000	0.0184	stratified	accuracy 95%	0.946	0.942	0.948	0.946
10,000	0.0046	naive	accuracy 95%	0.856	0.871	0.854	0.842
10,000	0.0046	stratified	accuracy 95%	0.840	0.834	0.836	0.832
10,000	0.0092	naive	accuracy 95%	0.940	0.942	0.942	0.936
10,000	0.0092	stratified	accuracy 95%	0.964	0.962	0.958	0.962
10,000	0.0184	naive	accuracy 95%	0.978	0.976	0.980	0.972
10,000	0.0184	stratified	accuracy 95%	0.964	0.970	0.960	0.966
50,000	0.0046	naive	accuracy 95%	0.888	0.888	0.890	0.890
50,000	0.0046	stratified	accuracy 95%	0.910	0.916	0.904	0.910
50,000	0.0092	naive	accuracy 95%	0.908	0.910	0.904	0.906
50,000	0.0092	stratified	accuracy 95%	0.916	0.910	0.910	0.900
50,000	0.0184	naive	accuracy 95%	0.902	0.896	0.892	0.896
50,000	0.0184	stratified	accuracy 95%	0.922	0.924	0.920	0.920
100,000	0.0046	naive	accuracy 95%	0.818	0.812	0.804	0.804
100,000	0.0046	stratified	accuracy 95%	0.838	0.838	0.832	0.828
100,000	0.0092	naive	accuracy 95%	0.848	0.850	0.844	0.846
100,000	0.0092	stratified	accuracy 95%	0.848	0.848	0.850	0.844
100,000	0.0184	naive	accuracy 95%	0.856	0.854	0.844	0.850
100,000	0.0184	stratified	accuracy 95%	0.860	0.860	0.846	0.852
1,000,000	0.0046	naive	accuracy 95%	0.396	0.398	0.394	0.396
1,000,000	0.0046	stratified	accuracy 95%	0.402	0.400	0.398	0.394
1,000,000	0.0092	naive	accuracy 95%	0.378	0.376	0.376	0.380
1,000,000	0.0092	stratified	accuracy 95%	0.424	0.422	0.428	0.426

Table B.8: Coverage of 95% confidence intervals for accuracy at the 95th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 99%	0.940	0.940	0.942	0.868
5,000	0.0046	stratified	accuracy 99%	0.942	0.996	0.942	0.878
5,000	0.0092	naive	accuracy 99%	0.974	0.996	0.976	0.976
5,000	0.0092	stratified	accuracy 99%	0.968	0.988	0.968	0.986
5,000	0.0184	naive	accuracy 99%	0.996	0.994	0.996	0.996
5,000	0.0184	stratified	accuracy 99%	1.000	1.000	0.996	1.000
10,000	0.0046	naive	accuracy 99%	0.846	0.961	0.842	0.836
10,000	0.0046	stratified	accuracy 99%	0.896	0.970	0.886	0.902
10,000	0.0092	naive	accuracy 99%	0.982	0.990	0.982	0.980
10,000	0.0092	stratified	accuracy 99%	0.976	0.978	0.974	0.980
10,000	0.0184	naive	accuracy 99%	0.996	0.996	0.996	0.996
10,000	0.0184	stratified	accuracy 99%	0.994	0.992	0.992	0.994
50,000	0.0046	naive	accuracy 99%	0.890	0.894	0.902	0.890
50,000	0.0046	stratified	accuracy 99%	0.880	0.880	0.888	0.882
50,000	0.0092	naive	accuracy 99%	0.916	0.928	0.922	0.926
50,000	0.0092	stratified	accuracy 99%	0.958	0.954	0.952	0.958
50,000	0.0184	naive	accuracy 99%	0.986	0.986	0.986	0.986
50,000	0.0184	stratified	accuracy 99%	0.990	0.988	0.992	0.986
100,000	0.0046	naive	accuracy 99%	0.856	0.856	0.854	0.856
100,000	0.0046	stratified	accuracy 99%	0.866	0.876	0.860	0.878
100,000	0.0092	naive	accuracy 99%	0.926	0.934	0.920	0.932
100,000	0.0092	stratified	accuracy 99%	0.896	0.898	0.890	0.894
100,000	0.0184	naive	accuracy 99%	0.984	0.984	0.982	0.982
100,000	0.0184	stratified	accuracy 99%	0.990	0.990	0.990	0.990
1,000,000	0.0046	naive	accuracy 99%	0.312	0.304	0.306	0.298
1,000,000	0.0046	stratified	accuracy 99%	0.350	0.346	0.346	0.344
1,000,000	0.0092	naive	accuracy 99%	0.408	0.412	0.408	0.400
1,000,000	0.0092	stratified	accuracy 99%	0.446	0.442	0.442	0.444

Table B.9: Coverage of 95% confidence intervals for accuracy at the 99th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 90%	0.734	0.797	0.726	0.735
5,000	0.0046	stratified	accuracy 90%	0.940	0.935	0.929	0.930
5,000	0.0092	naive	accuracy 90%	0.944	0.942	0.940	0.939
5,000	0.0092	stratified	accuracy 90%	0.939	0.938	0.935	0.932
5,000	0.0184	naive	accuracy 90%	0.946	0.947	0.944	0.943
5,000	0.0184	stratified	accuracy 90%	0.955	0.959	0.949	0.956
10,000	0.0046	naive	accuracy 90%	0.926	0.933	0.925	0.929
10,000	0.0046	stratified	accuracy 90%	0.940	0.942	0.938	0.936
10,000	0.0092	naive	accuracy 90%	0.952	0.954	0.949	0.948
10,000	0.0092	stratified	accuracy 90%	0.950	0.949	0.946	0.945
10,000	0.0184	naive	accuracy 90%	0.952	0.953	0.950	0.952
10,000	0.0184	stratified	accuracy 90%	0.940	0.938	0.933	0.933
50,000	0.0046	naive	accuracy 90%	0.924	0.918	0.918	0.917
50,000	0.0046	stratified	accuracy 90%	0.945	0.945	0.946	0.941
50,000	0.0092	naive	accuracy 90%	0.935	0.934	0.932	0.931
50,000	0.0092	stratified	accuracy 90%	0.938	0.940	0.937	0.937
50,000	0.0184	naive	accuracy 90%	0.942	0.937	0.935	0.935
50,000	0.0184	stratified	accuracy 90%	0.946	0.948	0.941	0.945
100,000	0.0046	naive	accuracy 90%	0.929	0.931	0.923	0.928
100,000	0.0046	stratified	accuracy 90%	0.933	0.938	0.929	0.935
100,000	0.0092	naive	accuracy 90%	0.946	0.942	0.941	0.937
100,000	0.0092	stratified	accuracy 90%	0.927	0.932	0.924	0.922
100,000	0.0184	naive	accuracy 90%	0.939	0.939	0.931	0.934
100,000	0.0184	stratified	accuracy 90%	0.933	0.935	0.926	0.930
1,000,000	0.0046	naive	accuracy 90%	0.827	0.824	0.818	0.820
1,000,000	0.0046	stratified	accuracy 90%	0.832	0.830	0.822	0.825
1,000,000	0.0092	naive	accuracy 90%	0.826	0.824	0.810	0.819
1,000,000	0.0092	stratified	accuracy 90%	0.843	0.844	0.835	0.839

Table B.10: Coverage of 95% confidence intervals for accuracy at the 90th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 95%	0.842	0.850	0.838	0.850
5,000	0.0046	stratified	accuracy 95%	0.938	0.935	0.937	0.935
5,000	0.0092	naive	accuracy 95%	0.959	0.959	0.954	0.957
5,000	0.0092	stratified	accuracy 95%	0.945	0.948	0.944	0.947
5,000	0.0184	naive	accuracy 95%	0.967	0.961	0.964	0.965
5,000	0.0184	stratified	accuracy 95%	0.956	0.955	0.951	0.950
10,000	0.0046	naive	accuracy 95%	0.935	0.932	0.926	0.921
10,000	0.0046	stratified	accuracy 95%	0.950	0.950	0.945	0.948
10,000	0.0092	naive	accuracy 95%	0.944	0.946	0.943	0.942
10,000	0.0092	stratified	accuracy 95%	0.950	0.953	0.949	0.950
10,000	0.0184	naive	accuracy 95%	0.945	0.946	0.941	0.937
10,000	0.0184	stratified	accuracy 95%	0.954	0.955	0.948	0.950
50,000	0.0046	naive	accuracy 95%	0.944	0.939	0.943	0.936
50,000	0.0046	stratified	accuracy 95%	0.944	0.943	0.944	0.942
50,000	0.0092	naive	accuracy 95%	0.952	0.950	0.947	0.944
50,000	0.0092	stratified	accuracy 95%	0.951	0.949	0.945	0.945
50,000	0.0184	naive	accuracy 95%	0.948	0.944	0.940	0.941
50,000	0.0184	stratified	accuracy 95%	0.943	0.940	0.946	0.941
100,000	0.0046	naive	accuracy 95%	0.939	0.934	0.932	0.935
100,000	0.0046	stratified	accuracy 95%	0.942	0.948	0.938	0.946
100,000	0.0092	naive	accuracy 95%	0.938	0.940	0.938	0.933
100,000	0.0092	stratified	accuracy 95%	0.924	0.927	0.928	0.927
100,000	0.0184	naive	accuracy 95%	0.947	0.948	0.939	0.944
100,000	0.0184	stratified	accuracy 95%	0.952	0.949	0.943	0.943
1,000,000	0.0046	naive	accuracy 95%	0.824	0.819	0.815	0.817
1,000,000	0.0046	stratified	accuracy 95%	0.852	0.854	0.849	0.856
1,000,000	0.0092	naive	accuracy 95%	0.850	0.849	0.844	0.847
1,000,000	0.0092	stratified	accuracy 95%	0.855	0.860	0.848	0.855

Table B.11: Coverage of 95% confidence intervals for accuracy at the 95th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 99%	0.938	0.865	0.934	0.821
5,000	0.0046	stratified	accuracy 99%	0.961	0.985	0.956	0.799
5,000	0.0092	naive	accuracy 99%	0.986	0.998	0.983	0.975
5,000	0.0092	stratified	accuracy 99%	0.977	0.992	0.971	0.984
5,000	0.0184	naive	accuracy 99%	0.994	0.997	0.993	0.994
5,000	0.0184	stratified	accuracy 99%	0.996	0.998	0.996	0.996
10,000	0.0046	naive	accuracy 99%	0.967	0.999	0.965	0.944
10,000	0.0046	stratified	accuracy 99%	0.967	0.997	0.958	0.975
10,000	0.0092	naive	accuracy 99%	0.985	0.991	0.985	0.991
10,000	0.0092	stratified	accuracy 99%	0.979	0.977	0.975	0.977
10,000	0.0184	naive	accuracy 99%	0.994	0.995	0.993	0.992
10,000	0.0184	stratified	accuracy 99%	0.999	0.999	0.998	0.998
50,000	0.0046	naive	accuracy 99%	0.977	0.974	0.971	0.970
50,000	0.0046	stratified	accuracy 99%	0.966	0.965	0.963	0.961
50,000	0.0092	naive	accuracy 99%	0.979	0.978	0.980	0.976
50,000	0.0092	stratified	accuracy 99%	0.979	0.980	0.979	0.975
50,000	0.0184	naive	accuracy 99%	0.996	0.996	0.996	0.995
50,000	0.0184	stratified	accuracy 99%	0.989	0.989	0.990	0.990
100,000	0.0046	naive	accuracy 99%	0.970	0.967	0.970	0.965
100,000	0.0046	stratified	accuracy 99%	0.973	0.972	0.972	0.970
100,000	0.0092	naive	accuracy 99%	0.980	0.978	0.976	0.974
100,000	0.0092	stratified	accuracy 99%	0.983	0.983	0.981	0.978
100,000	0.0184	naive	accuracy 99%	0.993	0.994	0.989	0.992
100,000	0.0184	stratified	accuracy 99%	0.989	0.989	0.987	0.988
1,000,000	0.0046	naive	accuracy 99%	0.877	0.880	0.872	0.873
1,000,000	0.0046	stratified	accuracy 99%	0.882	0.880	0.873	0.881
1,000,000	0.0092	naive	accuracy 99%	0.927	0.929	0.932	0.929
1,000,000	0.0092	stratified	accuracy 99%	0.925	0.918	0.919	0.918

Table B.12: Coverage of 95% confidence intervals for accuracy at the 99th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 90%	0.546	0.575	0.544	0.323
5,000	0.0046	stratified	accuracy 90%	0.549	0.548	0.549	0.547
5,000	0.0092	naive	accuracy 90%	0.282	0.284	0.281	0.290
5,000	0.0092	stratified	accuracy 90%	0.262	0.254	0.260	0.260
5,000	0.0184	naive	accuracy 90%	0.860	0.856	0.859	0.853
5,000	0.0184	stratified	accuracy 90%	0.855	0.858	0.852	0.858
10,000	0.0046	naive	accuracy 90%	0.158	0.151	0.160	0.155
10,000	0.0046	stratified	accuracy 90%	0.162	0.160	0.156	0.164
10,000	0.0092	naive	accuracy 90%	0.839	0.837	0.834	0.839
10,000	0.0092	stratified	accuracy 90%	0.844	0.849	0.838	0.840
10,000	0.0184	naive	accuracy 90%	0.935	0.934	0.936	0.936
10,000	0.0184	stratified	accuracy 90%	0.931	0.929	0.935	0.925
50,000	0.0046	naive	accuracy 90%	0.877	0.871	0.876	0.875
50,000	0.0046	stratified	accuracy 90%	0.895	0.898	0.887	0.889
50,000	0.0092	naive	accuracy 90%	0.931	0.932	0.928	0.930
50,000	0.0092	stratified	accuracy 90%	0.959	0.955	0.955	0.950
50,000	0.0184	naive	accuracy 90%	0.946	0.945	0.941	0.946
50,000	0.0184	stratified	accuracy 90%	0.931	0.933	0.928	0.927
100,000	0.0046	naive	accuracy 90%	0.915	0.906	0.911	0.904
100,000	0.0046	stratified	accuracy 90%	0.940	0.944	0.938	0.940
100,000	0.0092	naive	accuracy 90%	0.939	0.941	0.933	0.938
100,000	0.0092	stratified	accuracy 90%	0.922	0.922	0.915	0.915
100,000	0.0184	naive	accuracy 90%	0.934	0.933	0.931	0.926
100,000	0.0184	stratified	accuracy 90%	0.947	0.942	0.939	0.943
1,000,000	0.0046	naive	accuracy 90%	0.820	0.819	0.818	0.814
1,000,000	0.0046	stratified	accuracy 90%	0.847	0.843	0.843	0.841
1,000,000	0.0092	naive	accuracy 90%	0.815	0.821	0.813	0.820
1,000,000	0.0092	stratified	accuracy 90%	0.831	0.831	0.823	0.829

Table B.13: Coverage of 95% confidence intervals for accuracy at the 90th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 95%	0.547	0.585	0.543	0.343
5,000	0.0046	stratified	accuracy 95%	0.557	0.557	0.557	0.556
5,000	0.0092	naive	accuracy 95%	0.683	0.685	0.685	0.684
5,000	0.0092	stratified	accuracy 95%	0.679	0.679	0.681	0.682
5,000	0.0184	naive	accuracy 95%	0.915	0.916	0.911	0.909
5,000	0.0184	stratified	accuracy 95%	0.910	0.911	0.909	0.907
10,000	0.0046	naive	accuracy 95%	0.667	0.676	0.679	0.676
10,000	0.0046	stratified	accuracy 95%	0.693	0.692	0.692	0.687
10,000	0.0092	naive	accuracy 95%	0.935	0.937	0.928	0.934
10,000	0.0092	stratified	accuracy 95%	0.937	0.936	0.931	0.931
10,000	0.0184	naive	accuracy 95%	0.946	0.947	0.943	0.942
10,000	0.0184	stratified	accuracy 95%	0.931	0.935	0.926	0.933
50,000	0.0046	naive	accuracy 95%	0.938	0.936	0.932	0.934
50,000	0.0046	stratified	accuracy 95%	0.948	0.950	0.947	0.946
50,000	0.0092	naive	accuracy 95%	0.945	0.946	0.942	0.941
50,000	0.0092	stratified	accuracy 95%	0.956	0.954	0.951	0.953
50,000	0.0184	naive	accuracy 95%	0.952	0.948	0.947	0.942
50,000	0.0184	stratified	accuracy 95%	0.950	0.946	0.938	0.943
100,000	0.0046	naive	accuracy 95%	0.933	0.931	0.930	0.929
100,000	0.0046	stratified	accuracy 95%	0.937	0.935	0.932	0.934
100,000	0.0092	naive	accuracy 95%	0.927	0.920	0.919	0.910
100,000	0.0092	stratified	accuracy 95%	0.932	0.934	0.933	0.937
100,000	0.0184	naive	accuracy 95%	0.937	0.939	0.934	0.939
100,000	0.0184	stratified	accuracy 95%	0.943	0.947	0.942	0.940
1,000,000	0.0046	naive	accuracy 95%	0.819	0.820	0.812	0.818
1,000,000	0.0046	stratified	accuracy 95%	0.828	0.834	0.823	0.832
1,000,000	0.0092	naive	accuracy 95%	0.829	0.825	0.824	0.824
1,000,000	0.0092	stratified	accuracy 95%	0.832	0.835	0.830	0.830

Table B.14: Coverage of 95% confidence intervals for accuracy at the 95th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	accuracy 99%	0.762	0.748	0.755	0.691
5,000	0.0046	stratified	accuracy 99%	0.764	0.773	0.767	0.747
5,000	0.0092	naive	accuracy 99%	0.830	0.822	0.821	0.755
5,000	0.0092	stratified	accuracy 99%	0.845	0.847	0.833	0.821
5,000	0.0184	naive	accuracy 99%	0.964	0.966	0.960	0.959
5,000	0.0184	stratified	accuracy 99%	0.956	0.952	0.951	0.950
10,000	0.0046	naive	accuracy 99%	0.916	0.964	0.910	0.876
10,000	0.0046	stratified	accuracy 99%	0.902	0.944	0.900	0.893
10,000	0.0092	naive	accuracy 99%	0.951	0.961	0.946	0.948
10,000	0.0092	stratified	accuracy 99%	0.949	0.948	0.945	0.938
10,000	0.0184	naive	accuracy 99%	0.986	0.984	0.984	0.983
10,000	0.0184	stratified	accuracy 99%	0.986	0.985	0.981	0.982
50,000	0.0046	naive	accuracy 99%	0.942	0.942	0.938	0.937
50,000	0.0046	stratified	accuracy 99%	0.951	0.951	0.946	0.945
50,000	0.0092	naive	accuracy 99%	0.982	0.981	0.975	0.977
50,000	0.0092	stratified	accuracy 99%	0.978	0.979	0.969	0.976
50,000	0.0184	naive	accuracy 99%	0.993	0.992	0.990	0.992
50,000	0.0184	stratified	accuracy 99%	0.988	0.988	0.988	0.986
100,000	0.0046	naive	accuracy 99%	0.950	0.952	0.948	0.952
100,000	0.0046	stratified	accuracy 99%	0.960	0.954	0.950	0.948
100,000	0.0092	naive	accuracy 99%	0.973	0.968	0.971	0.969
100,000	0.0092	stratified	accuracy 99%	0.971	0.976	0.957	0.971
100,000	0.0184	naive	accuracy 99%	0.991	0.992	0.992	0.992
100,000	0.0184	stratified	accuracy 99%	0.993	0.993	0.992	0.991
1,000,000	0.0046	naive	accuracy 99%	0.879	0.875	0.876	0.870
1,000,000	0.0046	stratified	accuracy 99%	0.865	0.865	0.860	0.863
1,000,000	0.0092	naive	accuracy 99%	0.912	0.911	0.910	0.910
1,000,000	0.0092	stratified	accuracy 99%	0.922	0.921	0.920	0.917

Table B.15: Coverage of 95% confidence intervals for accuracy at the 99th percentile, logistic regression.

B.4 Sensitivity

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 90%	0.904	0.860	0.916	0.864
5,000	0.0046	stratified	sensitivity 90%	0.912	0.880	0.926	0.890
5,000	0.0092	naive	sensitivity 90%	0.926	0.904	0.930	0.912
5,000	0.0092	stratified	sensitivity 90%	0.938	0.938	0.938	0.942
5,000	0.0184	naive	sensitivity 90%	0.946	0.952	0.948	0.954
5,000	0.0184	stratified	sensitivity 90%	0.940	0.934	0.938	0.932
10,000	0.0046	naive	sensitivity 90%	0.916	0.922	0.918	0.932
10,000	0.0046	stratified	sensitivity 90%	0.940	0.942	0.944	0.942
10,000	0.0092	naive	sensitivity 90%	0.930	0.914	0.926	0.928
10,000	0.0092	stratified	sensitivity 90%	0.934	0.934	0.932	0.936
10,000	0.0184	naive	sensitivity 90%	0.918	0.912	0.916	0.910
10,000	0.0184	stratified	sensitivity 90%	0.934	0.946	0.940	0.946
50,000	0.0046	naive	sensitivity 90%	0.946	0.938	0.940	0.936
50,000	0.0046	stratified	sensitivity 90%	0.958	0.958	0.952	0.956
50,000	0.0092	naive	sensitivity 90%	0.938	0.950	0.932	0.944
50,000	0.0092	stratified	sensitivity 90%	0.924	0.934	0.922	0.930
50,000	0.0184	naive	sensitivity 90%	0.932	0.932	0.926	0.932
50,000	0.0184	stratified	sensitivity 90%	0.936	0.936	0.934	0.936
100,000	0.0046	naive	sensitivity 90%	0.932	0.942	0.924	0.926
100,000	0.0046	stratified	sensitivity 90%	0.928	0.920	0.922	0.914
100,000	0.0092	naive	sensitivity 90%	0.948	0.948	0.946	0.952
100,000	0.0092	stratified	sensitivity 90%	0.920	0.916	0.912	0.912
100,000	0.0184	naive	sensitivity 90%	0.898	0.896	0.888	0.890
100,000	0.0184	stratified	sensitivity 90%	0.926	0.932	0.926	0.932
1,000,000	0.0046	naive	sensitivity 90%	0.820	0.806	0.812	0.796
1,000,000	0.0046	stratified	sensitivity 90%	0.826	0.830	0.828	0.830
1,000,000	0.0092	naive	sensitivity 90%	0.794	0.804	0.790	0.794
1,000,000	0.0092	stratified	sensitivity 90%	0.782	0.774	0.764	0.774

Table B.16: Coverage of 95% confidence intervals for sensitivity at the 90th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 95%	0.886	0.862	0.896	0.862
5,000	0.0046	stratified	sensitivity 95%	0.910	0.882	0.926	0.888
5,000	0.0092	naive	sensitivity 95%	0.912	0.912	0.916	0.906
5,000	0.0092	stratified	sensitivity 95%	0.926	0.934	0.928	0.932
5,000	0.0184	naive	sensitivity 95%	0.940	0.938	0.948	0.946
5,000	0.0184	stratified	sensitivity 95%	0.944	0.942	0.938	0.932
10,000	0.0046	naive	sensitivity 95%	0.908	0.924	0.916	0.936
10,000	0.0046	stratified	sensitivity 95%	0.932	0.944	0.938	0.942
10,000	0.0092	naive	sensitivity 95%	0.926	0.924	0.922	0.920
10,000	0.0092	stratified	sensitivity 95%	0.928	0.934	0.942	0.936
10,000	0.0184	naive	sensitivity 95%	0.936	0.938	0.932	0.932
10,000	0.0184	stratified	sensitivity 95%	0.932	0.944	0.936	0.940
50,000	0.0046	naive	sensitivity 95%	0.940	0.932	0.940	0.932
50,000	0.0046	stratified	sensitivity 95%	0.936	0.932	0.936	0.936
50,000	0.0092	naive	sensitivity 95%	0.928	0.924	0.926	0.924
50,000	0.0092	stratified	sensitivity 95%	0.906	0.908	0.900	0.916
50,000	0.0184	naive	sensitivity 95%	0.902	0.912	0.900	0.906
50,000	0.0184	stratified	sensitivity 95%	0.900	0.900	0.890	0.898
100,000	0.0046	naive	sensitivity 95%	0.904	0.912	0.912	0.916
100,000	0.0046	stratified	sensitivity 95%	0.926	0.920	0.920	0.918
100,000	0.0092	naive	sensitivity 95%	0.954	0.960	0.964	0.956
100,000	0.0092	stratified	sensitivity 95%	0.934	0.928	0.934	0.926
100,000	0.0184	naive	sensitivity 95%	0.904	0.902	0.906	0.906
100,000	0.0184	stratified	sensitivity 95%	0.888	0.890	0.884	0.886
1,000,000	0.0046	naive	sensitivity 95%	0.784	0.780	0.774	0.780
1,000,000	0.0046	stratified	sensitivity 95%	0.828	0.828	0.822	0.824
1,000,000	0.0092	naive	sensitivity 95%	0.786	0.788	0.784	0.786
1,000,000	0.0092	stratified	sensitivity 95%	0.820	0.814	0.812	0.806

Table B.17: Coverage of 95% confidence intervals for sensitivity at the 95th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 99%	0.838	0.826	0.852	0.840
5,000	0.0046	stratified	sensitivity 99%	0.858	0.824	0.872	0.830
5,000	0.0092	naive	sensitivity 99%	0.896	0.880	0.910	0.886
5,000	0.0092	stratified	sensitivity 99%	0.866	0.878	0.880	0.886
5,000	0.0184	naive	sensitivity 99%	0.904	0.906	0.912	0.910
5,000	0.0184	stratified	sensitivity 99%	0.906	0.908	0.906	0.906
10,000	0.0046	naive	sensitivity 99%	0.886	0.876	0.894	0.888
10,000	0.0046	stratified	sensitivity 99%	0.872	0.880	0.888	0.894
10,000	0.0092	naive	sensitivity 99%	0.920	0.924	0.922	0.936
10,000	0.0092	stratified	sensitivity 99%	0.924	0.930	0.926	0.930
10,000	0.0184	naive	sensitivity 99%	0.886	0.916	0.898	0.924
10,000	0.0184	stratified	sensitivity 99%	0.914	0.918	0.924	0.918
50,000	0.0046	naive	sensitivity 99%	0.910	0.906	0.908	0.912
50,000	0.0046	stratified	sensitivity 99%	0.908	0.910	0.906	0.908
50,000	0.0092	naive	sensitivity 99%	0.922	0.930	0.926	0.930
50,000	0.0092	stratified	sensitivity 99%	0.910	0.922	0.914	0.918
50,000	0.0184	naive	sensitivity 99%	0.874	0.870	0.870	0.864
50,000	0.0184	stratified	sensitivity 99%	0.870	0.876	0.868	0.878
100,000	0.0046	naive	sensitivity 99%	0.922	0.918	0.924	0.910
100,000	0.0046	stratified	sensitivity 99%	0.926	0.926	0.926	0.924
100,000	0.0092	naive	sensitivity 99%	0.888	0.886	0.890	0.890
100,000	0.0092	stratified	sensitivity 99%	0.924	0.918	0.924	0.912
100,000	0.0184	naive	sensitivity 99%	0.898	0.900	0.902	0.906
100,000	0.0184	stratified	sensitivity 99%	0.850	0.848	0.850	0.844
1,000,000	0.0046	naive	sensitivity 99%	0.702	0.704	0.696	0.700
1,000,000	0.0046	stratified	sensitivity 99%	0.744	0.750	0.742	0.744
1,000,000	0.0092	naive	sensitivity 99%	0.662	0.674	0.652	0.652
1,000,000	0.0092	stratified	sensitivity 99%	0.660	0.670	0.656	0.660

Table B.18: Coverage of 95% confidence intervals for sensitivity at the 99th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 90%	0.889	0.896	0.899	0.894
5,000	0.0046	stratified	sensitivity 90%	0.926	0.909	0.934	0.913
5,000	0.0092	naive	sensitivity 90%	0.926	0.919	0.926	0.927
5,000	0.0092	stratified	sensitivity 90%	0.936	0.937	0.933	0.938
5,000	0.0184	naive	sensitivity 90%	0.929	0.938	0.934	0.940
5,000	0.0184	stratified	sensitivity 90%	0.945	0.942	0.944	0.943
10,000	0.0046	naive	sensitivity 90%	0.930	0.919	0.930	0.925
10,000	0.0046	stratified	sensitivity 90%	0.936	0.937	0.938	0.933
10,000	0.0092	naive	sensitivity 90%	0.939	0.935	0.935	0.934
10,000	0.0092	stratified	sensitivity 90%	0.931	0.926	0.934	0.930
10,000	0.0184	naive	sensitivity 90%	0.936	0.941	0.934	0.939
10,000	0.0184	stratified	sensitivity 90%	0.932	0.936	0.929	0.933
50,000	0.0046	naive	sensitivity 90%	0.932	0.933	0.927	0.934
50,000	0.0046	stratified	sensitivity 90%	0.944	0.943	0.944	0.938
50,000	0.0092	naive	sensitivity 90%	0.945	0.942	0.943	0.937
50,000	0.0092	stratified	sensitivity 90%	0.929	0.933	0.928	0.929
50,000	0.0184	naive	sensitivity 90%	0.935	0.935	0.929	0.933
50,000	0.0184	stratified	sensitivity 90%	0.939	0.937	0.936	0.935
100,000	0.0046	naive	sensitivity 90%	0.932	0.929	0.928	0.932
100,000	0.0046	stratified	sensitivity 90%	0.914	0.916	0.903	0.913
100,000	0.0092	naive	sensitivity 90%	0.940	0.938	0.936	0.938
100,000	0.0092	stratified	sensitivity 90%	0.925	0.923	0.918	0.924
100,000	0.0184	naive	sensitivity 90%	0.939	0.937	0.932	0.935
100,000	0.0184	stratified	sensitivity 90%	0.925	0.927	0.924	0.922
1,000,000	0.0046	naive	sensitivity 90%	0.833	0.828	0.836	0.819
1,000,000	0.0046	stratified	sensitivity 90%	0.815	0.817	0.815	0.809
1,000,000	0.0092	naive	sensitivity 90%	0.842	0.844	0.829	0.837
1,000,000	0.0092	stratified	sensitivity 90%	0.836	0.839	0.834	0.837

Table B.19: Coverage of 95% confidence intervals for sensitivity at the 90th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 95%	0.868	0.889	0.878	0.886
5,000	0.0046	stratified	sensitivity 95%	0.916	0.901	0.924	0.899
5,000	0.0092	naive	sensitivity 95%	0.938	0.925	0.937	0.930
5,000	0.0092	stratified	sensitivity 95%	0.930	0.930	0.927	0.934
5,000	0.0184	naive	sensitivity 95%	0.928	0.940	0.923	0.939
5,000	0.0184	stratified	sensitivity 95%	0.940	0.943	0.935	0.940
10,000	0.0046	naive	sensitivity 95%	0.914	0.915	0.922	0.922
10,000	0.0046	stratified	sensitivity 95%	0.931	0.938	0.941	0.940
10,000	0.0092	naive	sensitivity 95%	0.940	0.939	0.935	0.939
10,000	0.0092	stratified	sensitivity 95%	0.941	0.939	0.938	0.937
10,000	0.0184	naive	sensitivity 95%	0.930	0.923	0.932	0.923
10,000	0.0184	stratified	sensitivity 95%	0.942	0.939	0.940	0.934
50,000	0.0046	naive	sensitivity 95%	0.929	0.932	0.934	0.927
50,000	0.0046	stratified	sensitivity 95%	0.938	0.938	0.934	0.930
50,000	0.0092	naive	sensitivity 95%	0.938	0.940	0.937	0.937
50,000	0.0092	stratified	sensitivity 95%	0.935	0.934	0.935	0.935
50,000	0.0184	naive	sensitivity 95%	0.923	0.928	0.920	0.923
50,000	0.0184	stratified	sensitivity 95%	0.944	0.946	0.944	0.944
100,000	0.0046	naive	sensitivity 95%	0.936	0.936	0.929	0.936
100,000	0.0046	stratified	sensitivity 95%	0.916	0.919	0.912	0.918
100,000	0.0092	naive	sensitivity 95%	0.939	0.942	0.935	0.934
100,000	0.0092	stratified	sensitivity 95%	0.935	0.935	0.930	0.929
100,000	0.0184	naive	sensitivity 95%	0.942	0.940	0.942	0.929
100,000	0.0184	stratified	sensitivity 95%	0.931	0.933	0.926	0.931
1,000,000	0.0046	naive	sensitivity 95%	0.823	0.826	0.820	0.816
1,000,000	0.0046	stratified	sensitivity 95%	0.835	0.837	0.831	0.834
1,000,000	0.0092	naive	sensitivity 95%	0.845	0.843	0.842	0.839
1,000,000	0.0092	stratified	sensitivity 95%	0.854	0.856	0.851	0.857

Table B.20: Coverage of 95% confidence intervals for sensitivity at the 95th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 99%	0.814	0.804	0.834	0.813
5,000	0.0046	stratified	sensitivity 99%	0.839	0.824	0.863	0.830
5,000	0.0092	naive	sensitivity 99%	0.899	0.890	0.907	0.901
5,000	0.0092	stratified	sensitivity 99%	0.900	0.896	0.907	0.908
5,000	0.0184	naive	sensitivity 99%	0.936	0.943	0.933	0.945
5,000	0.0184	stratified	sensitivity 99%	0.919	0.922	0.930	0.933
10,000	0.0046	naive	sensitivity 99%	0.904	0.902	0.914	0.907
10,000	0.0046	stratified	sensitivity 99%	0.902	0.902	0.905	0.916
10,000	0.0092	naive	sensitivity 99%	0.917	0.913	0.915	0.915
10,000	0.0092	stratified	sensitivity 99%	0.926	0.926	0.928	0.929
10,000	0.0184	naive	sensitivity 99%	0.932	0.930	0.936	0.930
10,000	0.0184	stratified	sensitivity 99%	0.925	0.929	0.923	0.929
50,000	0.0046	naive	sensitivity 99%	0.938	0.938	0.937	0.937
50,000	0.0046	stratified	sensitivity 99%	0.934	0.928	0.934	0.931
50,000	0.0092	naive	sensitivity 99%	0.939	0.940	0.940	0.934
50,000	0.0092	stratified	sensitivity 99%	0.926	0.922	0.927	0.923
50,000	0.0184	naive	sensitivity 99%	0.951	0.947	0.945	0.947
50,000	0.0184	stratified	sensitivity 99%	0.947	0.949	0.942	0.945
100,000	0.0046	naive	sensitivity 99%	0.933	0.928	0.932	0.931
100,000	0.0046	stratified	sensitivity 99%	0.918	0.922	0.919	0.921
100,000	0.0092	naive	sensitivity 99%	0.923	0.926	0.922	0.923
100,000	0.0092	stratified	sensitivity 99%	0.936	0.928	0.927	0.928
100,000	0.0184	naive	sensitivity 99%	0.932	0.932	0.933	0.929
100,000	0.0184	stratified	sensitivity 99%	0.935	0.933	0.930	0.930
1,000,000	0.0046	naive	sensitivity 99%	0.818	0.816	0.815	0.812
1,000,000	0.0046	stratified	sensitivity 99%	0.839	0.827	0.835	0.826
1,000,000	0.0092	naive	sensitivity 99%	0.845	0.850	0.845	0.846
1,000,000	0.0092	stratified	sensitivity 99%	0.847	0.847	0.845	0.846

Table B.21: Coverage of 95% confidence intervals for sensitivity at the 99th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 90%	0.827	0.828	0.849	0.826
5,000	0.0046	stratified	sensitivity 90%	0.835	0.825	0.857	0.829
5,000	0.0092	naive	sensitivity 90%	0.786	0.805	0.805	0.824
5,000	0.0092	stratified	sensitivity 90%	0.834	0.844	0.844	0.857
5,000	0.0184	naive	sensitivity 90%	0.842	0.855	0.839	0.864
5,000	0.0184	stratified	sensitivity 90%	0.837	0.842	0.837	0.841
10,000	0.0046	naive	sensitivity 90%	0.795	0.815	0.812	0.827
10,000	0.0046	stratified	sensitivity 90%	0.840	0.850	0.851	0.855
10,000	0.0092	naive	sensitivity 90%	0.834	0.860	0.842	0.864
10,000	0.0092	stratified	sensitivity 90%	0.854	0.857	0.856	0.856
10,000	0.0184	naive	sensitivity 90%	0.879	0.887	0.873	0.890
10,000	0.0184	stratified	sensitivity 90%	0.879	0.884	0.877	0.886
50,000	0.0046	naive	sensitivity 90%	0.901	0.902	0.892	0.905
50,000	0.0046	stratified	sensitivity 90%	0.901	0.902	0.898	0.903
50,000	0.0092	naive	sensitivity 90%	0.907	0.910	0.910	0.904
50,000	0.0092	stratified	sensitivity 90%	0.928	0.934	0.921	0.920
50,000	0.0184	naive	sensitivity 90%	0.929	0.929	0.926	0.921
50,000	0.0184	stratified	sensitivity 90%	0.933	0.937	0.931	0.931
100,000	0.0046	naive	sensitivity 90%	0.914	0.917	0.910	0.910
100,000	0.0046	stratified	sensitivity 90%	0.909	0.911	0.911	0.910
100,000	0.0092	naive	sensitivity 90%	0.936	0.933	0.932	0.930
100,000	0.0092	stratified	sensitivity 90%	0.927	0.926	0.921	0.919
100,000	0.0184	naive	sensitivity 90%	0.936	0.939	0.932	0.936
100,000	0.0184	stratified	sensitivity 90%	0.940	0.941	0.934	0.937
1,000,000	0.0046	naive	sensitivity 90%	0.838	0.833	0.832	0.832
1,000,000	0.0046	stratified	sensitivity 90%	0.821	0.814	0.809	0.813
1,000,000	0.0092	naive	sensitivity 90%	0.830	0.832	0.821	0.816
1,000,000	0.0092	stratified	sensitivity 90%	0.836	0.835	0.827	0.834

Table B.22: Coverage of 95% confidence intervals for sensitivity at the 90th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 95%	0.882	0.863	0.909	0.873
5,000	0.0046	stratified	sensitivity 95%	0.879	0.872	0.897	0.858
5,000	0.0092	naive	sensitivity 95%	0.832	0.842	0.849	0.855
5,000	0.0092	stratified	sensitivity 95%	0.854	0.853	0.866	0.874
5,000	0.0184	naive	sensitivity 95%	0.859	0.873	0.871	0.878
5,000	0.0184	stratified	sensitivity 95%	0.868	0.873	0.874	0.883
10,000	0.0046	naive	sensitivity 95%	0.832	0.853	0.856	0.861
10,000	0.0046	stratified	sensitivity 95%	0.847	0.857	0.867	0.874
10,000	0.0092	naive	sensitivity 95%	0.859	0.881	0.863	0.892
10,000	0.0092	stratified	sensitivity 95%	0.868	0.875	0.879	0.885
10,000	0.0184	naive	sensitivity 95%	0.904	0.902	0.904	0.903
10,000	0.0184	stratified	sensitivity 95%	0.917	0.923	0.917	0.923
50,000	0.0046	naive	sensitivity 95%	0.921	0.917	0.922	0.915
50,000	0.0046	stratified	sensitivity 95%	0.901	0.901	0.897	0.900
50,000	0.0092	naive	sensitivity 95%	0.926	0.921	0.921	0.917
50,000	0.0092	stratified	sensitivity 95%	0.938	0.936	0.940	0.936
50,000	0.0184	naive	sensitivity 95%	0.939	0.943	0.936	0.940
50,000	0.0184	stratified	sensitivity 95%	0.925	0.930	0.925	0.927
100,000	0.0046	naive	sensitivity 95%	0.897	0.904	0.901	0.901
100,000	0.0046	stratified	sensitivity 95%	0.910	0.911	0.907	0.910
100,000	0.0092	naive	sensitivity 95%	0.920	0.922	0.922	0.916
100,000	0.0092	stratified	sensitivity 95%	0.918	0.920	0.915	0.919
100,000	0.0184	naive	sensitivity 95%	0.924	0.927	0.922	0.924
100,000	0.0184	stratified	sensitivity 95%	0.948	0.947	0.944	0.942
1,000,000	0.0046	naive	sensitivity 95%	0.827	0.827	0.830	0.823
1,000,000	0.0046	stratified	sensitivity 95%	0.815	0.815	0.814	0.817
1,000,000	0.0092	naive	sensitivity 95%	0.839	0.833	0.836	0.827
1,000,000	0.0092	stratified	sensitivity 95%	0.841	0.839	0.837	0.836

Table B.23: Coverage of 95% confidence intervals for sensitivity at the 95th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	sensitivity 99%	0.869	0.860	0.877	0.862
5,000	0.0046	stratified	sensitivity 99%	0.865	0.853	0.862	0.852
5,000	0.0092	naive	sensitivity 99%	0.864	0.863	0.887	0.866
5,000	0.0092	stratified	sensitivity 99%	0.887	0.885	0.894	0.904
5,000	0.0184	naive	sensitivity 99%	0.916	0.907	0.925	0.919
5,000	0.0184	stratified	sensitivity 99%	0.913	0.917	0.917	0.925
10,000	0.0046	naive	sensitivity 99%	0.853	0.837	0.883	0.866
10,000	0.0046	stratified	sensitivity 99%	0.880	0.880	0.893	0.892
10,000	0.0092	naive	sensitivity 99%	0.910	0.916	0.922	0.922
10,000	0.0092	stratified	sensitivity 99%	0.894	0.897	0.898	0.896
10,000	0.0184	naive	sensitivity 99%	0.940	0.932	0.940	0.935
10,000	0.0184	stratified	sensitivity 99%	0.921	0.924	0.925	0.924
50,000	0.0046	naive	sensitivity 99%	0.928	0.935	0.925	0.929
50,000	0.0046	stratified	sensitivity 99%	0.916	0.912	0.915	0.911
50,000	0.0092	naive	sensitivity 99%	0.929	0.922	0.927	0.922
50,000	0.0092	stratified	sensitivity 99%	0.918	0.928	0.923	0.923
50,000	0.0184	naive	sensitivity 99%	0.939	0.946	0.939	0.946
50,000	0.0184	stratified	sensitivity 99%	0.943	0.946	0.937	0.939
100,000	0.0046	naive	sensitivity 99%	0.914	0.917	0.915	0.911
100,000	0.0046	stratified	sensitivity 99%	0.922	0.922	0.921	0.921
100,000	0.0092	naive	sensitivity 99%	0.929	0.929	0.923	0.925
100,000	0.0092	stratified	sensitivity 99%	0.924	0.927	0.923	0.918
100,000	0.0184	naive	sensitivity 99%	0.949	0.947	0.946	0.948
100,000	0.0184	stratified	sensitivity 99%	0.942	0.942	0.940	0.941
1,000,000	0.0046	naive	sensitivity 99%	0.803	0.805	0.802	0.805
1,000,000	0.0046	stratified	sensitivity 99%	0.839	0.833	0.828	0.824
1,000,000	0.0092	naive	sensitivity 99%	0.838	0.845	0.829	0.841
1,000,000	0.0092	stratified	sensitivity 99%	0.830	0.828	0.831	0.827

Table B.24: Coverage of 95% confidence intervals for sensitivity at the 99th percentile, logistic regression.

B.5 Specificity

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 90%	0.890	0.888	0.892	0.880
5,000	0.0046	stratified	specificity 90%	0.916	0.922	0.914	0.920
5,000	0.0092	naive	specificity 90%	0.922	0.924	0.916	0.924
5,000	0.0092	stratified	specificity 90%	0.934	0.940	0.916	0.922
5,000	0.0184	naive	specificity 90%	0.916	0.914	0.908	0.914
5,000	0.0184	stratified	specificity 90%	0.946	0.940	0.938	0.942
10,000	0.0046	naive	specificity 90%	0.872	0.868	0.866	0.872
10,000	0.0046	stratified	specificity 90%	0.820	0.824	0.816	0.824
10,000	0.0092	naive	specificity 90%	0.928	0.922	0.918	0.924
10,000	0.0092	stratified	specificity 90%	0.900	0.910	0.900	0.912
10,000	0.0184	naive	specificity 90%	0.946	0.946	0.940	0.944
10,000	0.0184	stratified	specificity 90%	0.930	0.932	0.928	0.928
50,000	0.0046	naive	specificity 90%	0.838	0.844	0.826	0.836
50,000	0.0046	stratified	specificity 90%	0.824	0.832	0.832	0.832
50,000	0.0092	naive	specificity 90%	0.840	0.848	0.842	0.840
50,000	0.0092	stratified	specificity 90%	0.868	0.868	0.856	0.856
50,000	0.0184	naive	specificity 90%	0.876	0.888	0.880	0.870
50,000	0.0184	stratified	specificity 90%	0.878	0.874	0.872	0.882
100,000	0.0046	naive	specificity 90%	0.784	0.786	0.770	0.776
100,000	0.0046	stratified	specificity 90%	0.762	0.760	0.764	0.756
100,000	0.0092	naive	specificity 90%	0.806	0.810	0.808	0.796
100,000	0.0092	stratified	specificity 90%	0.810	0.798	0.800	0.792
100,000	0.0184	naive	specificity 90%	0.792	0.782	0.788	0.780
100,000	0.0184	stratified	specificity 90%	0.824	0.822	0.810	0.814
1,000,000	0.0046	naive	specificity 90%	0.374	0.368	0.358	0.368
1,000,000	0.0046	stratified	specificity 90%	0.330	0.330	0.326	0.328
1,000,000	0.0092	naive	specificity 90%	0.302	0.306	0.302	0.306
1,000,000	0.0092	stratified	specificity 90%	0.300	0.306	0.298	0.300

Table B.25: Coverage of 95% confidence intervals for specificity at the 90th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 95%	0.906	0.908	0.906	0.908
5,000	0.0046	stratified	specificity 95%	0.924	0.922	0.918	0.914
5,000	0.0092	naive	specificity 95%	0.942	0.946	0.932	0.936
5,000	0.0092	stratified	specificity 95%	0.932	0.932	0.930	0.928
5,000	0.0184	naive	specificity 95%	0.912	0.918	0.916	0.910
5,000	0.0184	stratified	specificity 95%	0.944	0.940	0.938	0.938
10,000	0.0046	naive	specificity 95%	0.854	0.844	0.852	0.844
10,000	0.0046	stratified	specificity 95%	0.838	0.838	0.830	0.832
10,000	0.0092	naive	specificity 95%	0.946	0.942	0.946	0.934
10,000	0.0092	stratified	specificity 95%	0.956	0.958	0.954	0.958
10,000	0.0184	naive	specificity 95%	0.962	0.958	0.952	0.960
10,000	0.0184	stratified	specificity 95%	0.952	0.948	0.948	0.944
50,000	0.0046	naive	specificity 95%	0.878	0.882	0.872	0.882
50,000	0.0046	stratified	specificity 95%	0.900	0.896	0.896	0.898
50,000	0.0092	naive	specificity 95%	0.874	0.880	0.876	0.874
50,000	0.0092	stratified	specificity 95%	0.890	0.882	0.888	0.874
50,000	0.0184	naive	specificity 95%	0.854	0.852	0.840	0.850
50,000	0.0184	stratified	specificity 95%	0.876	0.884	0.870	0.884
100,000	0.0046	naive	specificity 95%	0.788	0.786	0.786	0.782
100,000	0.0046	stratified	specificity 95%	0.820	0.826	0.816	0.822
100,000	0.0092	naive	specificity 95%	0.816	0.822	0.810	0.818
100,000	0.0092	stratified	specificity 95%	0.828	0.832	0.828	0.828
100,000	0.0184	naive	specificity 95%	0.826	0.818	0.818	0.810
100,000	0.0184	stratified	specificity 95%	0.816	0.810	0.802	0.806
1,000,000	0.0046	naive	specificity 95%	0.382	0.382	0.380	0.374
1,000,000	0.0046	stratified	specificity 95%	0.398	0.388	0.396	0.378
1,000,000	0.0092	naive	specificity 95%	0.354	0.354	0.348	0.350
1,000,000	0.0092	stratified	specificity 95%	0.394	0.400	0.394	0.398

Table B.26: Coverage of 95% confidence intervals for specificity at the 95th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 99%	0.890	0.888	0.892	0.896
5,000	0.0046	stratified	specificity 99%	0.898	0.898	0.898	0.900
5,000	0.0092	naive	specificity 99%	0.914	0.902	0.910	0.910
5,000	0.0092	stratified	specificity 99%	0.922	0.922	0.914	0.920
5,000	0.0184	naive	specificity 99%	0.938	0.936	0.944	0.936
5,000	0.0184	stratified	specificity 99%	0.948	0.950	0.952	0.948
10,000	0.0046	naive	specificity 99%	0.778	0.784	0.768	0.776
10,000	0.0046	stratified	specificity 99%	0.860	0.856	0.850	0.844
10,000	0.0092	naive	specificity 99%	0.922	0.920	0.916	0.912
10,000	0.0092	stratified	specificity 99%	0.928	0.934	0.926	0.926
10,000	0.0184	naive	specificity 99%	0.934	0.934	0.932	0.932
10,000	0.0184	stratified	specificity 99%	0.918	0.920	0.916	0.918
50,000	0.0046	naive	specificity 99%	0.806	0.800	0.804	0.788
50,000	0.0046	stratified	specificity 99%	0.820	0.802	0.818	0.802
50,000	0.0092	naive	specificity 99%	0.804	0.796	0.808	0.796
50,000	0.0092	stratified	specificity 99%	0.810	0.804	0.810	0.796
50,000	0.0184	naive	specificity 99%	0.810	0.812	0.820	0.806
50,000	0.0184	stratified	specificity 99%	0.824	0.802	0.820	0.810
100,000	0.0046	naive	specificity 99%	0.786	0.796	0.790	0.798
100,000	0.0046	stratified	specificity 99%	0.748	0.750	0.760	0.748
100,000	0.0092	naive	specificity 99%	0.768	0.762	0.762	0.762
100,000	0.0092	stratified	specificity 99%	0.732	0.738	0.734	0.734
100,000	0.0184	naive	specificity 99%	0.768	0.774	0.776	0.776
100,000	0.0184	stratified	specificity 99%	0.788	0.790	0.792	0.790
1,000,000	0.0046	naive	specificity 99%	0.250	0.252	0.246	0.244
1,000,000	0.0046	stratified	specificity 99%	0.290	0.290	0.284	0.282
1,000,000	0.0092	naive	specificity 99%	0.258	0.260	0.258	0.256
1,000,000	0.0092	stratified	specificity 99%	0.270	0.274	0.270	0.272

Table B.27: Coverage of 95% confidence intervals for specificity at the 99th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 90%	0.728	0.712	0.717	0.713
5,000	0.0046	stratified	specificity 90%	0.939	0.938	0.934	0.931
5,000	0.0092	naive	specificity 90%	0.943	0.942	0.942	0.939
5,000	0.0092	stratified	specificity 90%	0.937	0.931	0.934	0.935
5,000	0.0184	naive	specificity 90%	0.940	0.947	0.941	0.943
5,000	0.0184	stratified	specificity 90%	0.949	0.946	0.941	0.942
10,000	0.0046	naive	specificity 90%	0.931	0.930	0.928	0.928
10,000	0.0046	stratified	specificity 90%	0.942	0.939	0.937	0.930
10,000	0.0092	naive	specificity 90%	0.948	0.951	0.950	0.949
10,000	0.0092	stratified	specificity 90%	0.949	0.954	0.948	0.947
10,000	0.0184	naive	specificity 90%	0.956	0.954	0.944	0.952
10,000	0.0184	stratified	specificity 90%	0.943	0.942	0.939	0.939
50,000	0.0046	naive	specificity 90%	0.913	0.928	0.914	0.917
50,000	0.0046	stratified	specificity 90%	0.947	0.946	0.942	0.945
50,000	0.0092	naive	specificity 90%	0.940	0.938	0.936	0.936
50,000	0.0092	stratified	specificity 90%	0.934	0.937	0.934	0.936
50,000	0.0184	naive	specificity 90%	0.936	0.932	0.936	0.930
50,000	0.0184	stratified	specificity 90%	0.943	0.944	0.939	0.944
100,000	0.0046	naive	specificity 90%	0.921	0.929	0.919	0.925
100,000	0.0046	stratified	specificity 90%	0.942	0.939	0.935	0.934
100,000	0.0092	naive	specificity 90%	0.943	0.941	0.940	0.936
100,000	0.0092	stratified	specificity 90%	0.926	0.924	0.926	0.922
100,000	0.0184	naive	specificity 90%	0.927	0.929	0.918	0.923
100,000	0.0184	stratified	specificity 90%	0.931	0.933	0.929	0.928
1,000,000	0.0046	naive	specificity 90%	0.824	0.827	0.815	0.819
1,000,000	0.0046	stratified	specificity 90%	0.822	0.826	0.815	0.822
1,000,000	0.0092	naive	specificity 90%	0.815	0.814	0.812	0.815
1,000,000	0.0092	stratified	specificity 90%	0.833	0.829	0.830	0.826

Table B.28: Coverage of 95% confidence intervals for specificity at the 90th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 95%	0.824	0.825	0.816	0.818
5,000	0.0046	stratified	specificity 95%	0.940	0.944	0.939	0.942
5,000	0.0092	naive	specificity 95%	0.945	0.947	0.945	0.952
5,000	0.0092	stratified	specificity 95%	0.943	0.948	0.940	0.937
5,000	0.0184	naive	specificity 95%	0.959	0.959	0.952	0.954
5,000	0.0184	stratified	specificity 95%	0.945	0.945	0.943	0.941
10,000	0.0046	naive	specificity 95%	0.935	0.929	0.924	0.926
10,000	0.0046	stratified	specificity 95%	0.949	0.942	0.941	0.940
10,000	0.0092	naive	specificity 95%	0.938	0.939	0.942	0.938
10,000	0.0092	stratified	specificity 95%	0.944	0.938	0.939	0.934
10,000	0.0184	naive	specificity 95%	0.937	0.936	0.931	0.935
10,000	0.0184	stratified	specificity 95%	0.944	0.936	0.932	0.930
50,000	0.0046	naive	specificity 95%	0.942	0.941	0.941	0.940
50,000	0.0046	stratified	specificity 95%	0.941	0.938	0.930	0.935
50,000	0.0092	naive	specificity 95%	0.941	0.944	0.938	0.938
50,000	0.0092	stratified	specificity 95%	0.951	0.952	0.948	0.949
50,000	0.0184	naive	specificity 95%	0.938	0.944	0.938	0.937
50,000	0.0184	stratified	specificity 95%	0.940	0.934	0.935	0.928
100,000	0.0046	naive	specificity 95%	0.938	0.936	0.932	0.934
100,000	0.0046	stratified	specificity 95%	0.941	0.945	0.941	0.939
100,000	0.0092	naive	specificity 95%	0.943	0.943	0.934	0.940
100,000	0.0092	stratified	specificity 95%	0.923	0.916	0.918	0.918
100,000	0.0184	naive	specificity 95%	0.937	0.941	0.935	0.936
100,000	0.0184	stratified	specificity 95%	0.940	0.938	0.936	0.938
1,000,000	0.0046	naive	specificity 95%	0.822	0.820	0.817	0.815
1,000,000	0.0046	stratified	specificity 95%	0.853	0.854	0.847	0.851
1,000,000	0.0092	naive	specificity 95%	0.845	0.836	0.837	0.840
1,000,000	0.0092	stratified	specificity 95%	0.854	0.857	0.852	0.857

Table B.29: Coverage of 95% confidence intervals for specificity at the 95th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 99%	0.913	0.921	0.911	0.914
5,000	0.0046	stratified	specificity 99%	0.939	0.944	0.934	0.931
5,000	0.0092	naive	specificity 99%	0.943	0.947	0.946	0.942
5,000	0.0092	stratified	specificity 99%	0.947	0.948	0.944	0.943
5,000	0.0184	naive	specificity 99%	0.951	0.947	0.950	0.948
5,000	0.0184	stratified	specificity 99%	0.953	0.956	0.946	0.955
10,000	0.0046	naive	specificity 99%	0.937	0.939	0.932	0.939
10,000	0.0046	stratified	specificity 99%	0.943	0.943	0.940	0.938
10,000	0.0092	naive	specificity 99%	0.945	0.948	0.944	0.949
10,000	0.0092	stratified	specificity 99%	0.942	0.944	0.942	0.936
10,000	0.0184	naive	specificity 99%	0.943	0.943	0.940	0.945
10,000	0.0184	stratified	specificity 99%	0.939	0.944	0.937	0.940
50,000	0.0046	naive	specificity 99%	0.943	0.945	0.946	0.941
50,000	0.0046	stratified	specificity 99%	0.942	0.939	0.939	0.936
50,000	0.0092	naive	specificity 99%	0.939	0.946	0.936	0.940
50,000	0.0092	stratified	specificity 99%	0.942	0.945	0.937	0.941
50,000	0.0184	naive	specificity 99%	0.937	0.941	0.936	0.939
50,000	0.0184	stratified	specificity 99%	0.929	0.931	0.926	0.932
100,000	0.0046	naive	specificity 99%	0.945	0.946	0.945	0.941
100,000	0.0046	stratified	specificity 99%	0.945	0.942	0.943	0.942
100,000	0.0092	naive	specificity 99%	0.936	0.935	0.931	0.929
100,000	0.0092	stratified	specificity 99%	0.934	0.939	0.928	0.933
100,000	0.0184	naive	specificity 99%	0.924	0.936	0.920	0.926
100,000	0.0184	stratified	specificity 99%	0.939	0.934	0.933	0.934
1,000,000	0.0046	naive	specificity 99%	0.830	0.830	0.827	0.827
1,000,000	0.0046	stratified	specificity 99%	0.828	0.827	0.827	0.828
1,000,000	0.0092	naive	specificity 99%	0.845	0.847	0.837	0.846
1,000,000	0.0092	stratified	specificity 99%	0.851	0.844	0.839	0.837

Table B.30: Coverage of 95% confidence intervals for specificity at the 99th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 90%	0.546	0.544	0.546	0.544
5,000	0.0046	stratified	specificity 90%	0.549	0.547	0.549	0.547
5,000	0.0092	naive	specificity 90%	0.241	0.241	0.246	0.237
5,000	0.0092	stratified	specificity 90%	0.225	0.223	0.220	0.220
5,000	0.0184	naive	specificity 90%	0.832	0.832	0.833	0.827
5,000	0.0184	stratified	specificity 90%	0.831	0.836	0.830	0.836
10,000	0.0046	naive	specificity 90%	0.129	0.129	0.127	0.129
10,000	0.0046	stratified	specificity 90%	0.141	0.143	0.140	0.143
10,000	0.0092	naive	specificity 90%	0.815	0.812	0.809	0.805
10,000	0.0092	stratified	specificity 90%	0.828	0.829	0.823	0.822
10,000	0.0184	naive	specificity 90%	0.932	0.930	0.926	0.928
10,000	0.0184	stratified	specificity 90%	0.941	0.942	0.940	0.935
50,000	0.0046	naive	specificity 90%	0.868	0.866	0.870	0.861
50,000	0.0046	stratified	specificity 90%	0.888	0.886	0.876	0.884
50,000	0.0092	naive	specificity 90%	0.922	0.923	0.920	0.921
50,000	0.0092	stratified	specificity 90%	0.956	0.954	0.951	0.952
50,000	0.0184	naive	specificity 90%	0.942	0.941	0.935	0.939
50,000	0.0184	stratified	specificity 90%	0.931	0.936	0.933	0.927
100,000	0.0046	naive	specificity 90%	0.910	0.909	0.906	0.908
100,000	0.0046	stratified	specificity 90%	0.942	0.940	0.934	0.935
100,000	0.0092	naive	specificity 90%	0.936	0.941	0.931	0.940
100,000	0.0092	stratified	specificity 90%	0.919	0.923	0.915	0.917
100,000	0.0184	naive	specificity 90%	0.929	0.922	0.926	0.923
100,000	0.0184	stratified	specificity 90%	0.943	0.941	0.936	0.938
1,000,000	0.0046	naive	specificity 90%	0.812	0.812	0.812	0.809
1,000,000	0.0046	stratified	specificity 90%	0.845	0.840	0.841	0.830
1,000,000	0.0092	naive	specificity 90%	0.818	0.815	0.815	0.810
1,000,000	0.0092	stratified	specificity 90%	0.819	0.826	0.819	0.822

Table B.31: Coverage of 95% confidence intervals for specificity at the 90th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 95%	0.539	0.542	0.540	0.543
5,000	0.0046	stratified	specificity 95%	0.554	0.554	0.555	0.553
5,000	0.0092	naive	specificity 95%	0.638	0.634	0.641	0.634
5,000	0.0092	stratified	specificity 95%	0.645	0.648	0.647	0.642
5,000	0.0184	naive	specificity 95%	0.902	0.900	0.894	0.894
5,000	0.0184	stratified	specificity 95%	0.897	0.897	0.889	0.888
10,000	0.0046	naive	specificity 95%	0.637	0.637	0.640	0.639
10,000	0.0046	stratified	specificity 95%	0.662	0.660	0.660	0.660
10,000	0.0092	naive	specificity 95%	0.927	0.925	0.920	0.924
10,000	0.0092	stratified	specificity 95%	0.926	0.924	0.925	0.920
10,000	0.0184	naive	specificity 95%	0.945	0.938	0.938	0.937
10,000	0.0184	stratified	specificity 95%	0.922	0.921	0.911	0.917
50,000	0.0046	naive	specificity 95%	0.935	0.928	0.930	0.931
50,000	0.0046	stratified	specificity 95%	0.946	0.947	0.942	0.945
50,000	0.0092	naive	specificity 95%	0.936	0.935	0.928	0.934
50,000	0.0092	stratified	specificity 95%	0.952	0.948	0.943	0.950
50,000	0.0184	naive	specificity 95%	0.943	0.945	0.940	0.941
50,000	0.0184	stratified	specificity 95%	0.947	0.944	0.938	0.947
100,000	0.0046	naive	specificity 95%	0.934	0.933	0.931	0.928
100,000	0.0046	stratified	specificity 95%	0.936	0.937	0.929	0.932
100,000	0.0092	naive	specificity 95%	0.926	0.922	0.919	0.923
100,000	0.0092	stratified	specificity 95%	0.932	0.932	0.934	0.930
100,000	0.0184	naive	specificity 95%	0.922	0.925	0.921	0.921
100,000	0.0184	stratified	specificity 95%	0.935	0.932	0.934	0.930
1,000,000	0.0046	naive	specificity 95%	0.817	0.823	0.814	0.823
1,000,000	0.0046	stratified	specificity 95%	0.822	0.824	0.814	0.819
1,000,000	0.0092	naive	specificity 95%	0.820	0.820	0.814	0.808
1,000,000	0.0092	stratified	specificity 95%	0.818	0.821	0.812	0.822

Table B.32: Coverage of 95% confidence intervals for specificity at the 95th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	specificity 99%	0.732	0.729	0.726	0.725
5,000	0.0046	stratified	specificity 99%	0.736	0.739	0.732	0.728
5,000	0.0092	naive	specificity 99%	0.750	0.750	0.737	0.736
5,000	0.0092	stratified	specificity 99%	0.768	0.769	0.750	0.761
5,000	0.0184	naive	specificity 99%	0.876	0.879	0.865	0.868
5,000	0.0184	stratified	specificity 99%	0.861	0.861	0.839	0.841
10,000	0.0046	naive	specificity 99%	0.892	0.886	0.878	0.877
10,000	0.0046	stratified	specificity 99%	0.860	0.861	0.851	0.854
10,000	0.0092	naive	specificity 99%	0.900	0.898	0.882	0.888
10,000	0.0092	stratified	specificity 99%	0.873	0.875	0.863	0.864
10,000	0.0184	naive	specificity 99%	0.909	0.904	0.899	0.896
10,000	0.0184	stratified	specificity 99%	0.907	0.910	0.896	0.893
50,000	0.0046	naive	specificity 99%	0.918	0.915	0.915	0.910
50,000	0.0046	stratified	specificity 99%	0.911	0.911	0.917	0.910
50,000	0.0092	naive	specificity 99%	0.940	0.939	0.933	0.931
50,000	0.0092	stratified	specificity 99%	0.928	0.926	0.920	0.927
50,000	0.0184	naive	specificity 99%	0.948	0.950	0.943	0.947
50,000	0.0184	stratified	specificity 99%	0.911	0.912	0.906	0.910
100,000	0.0046	naive	specificity 99%	0.918	0.914	0.921	0.914
100,000	0.0046	stratified	specificity 99%	0.927	0.924	0.922	0.921
100,000	0.0092	naive	specificity 99%	0.931	0.935	0.932	0.935
100,000	0.0092	stratified	specificity 99%	0.921	0.928	0.917	0.919
100,000	0.0184	naive	specificity 99%	0.932	0.935	0.926	0.931
100,000	0.0184	stratified	specificity 99%	0.942	0.944	0.939	0.939
1,000,000	0.0046	naive	specificity 99%	0.845	0.844	0.836	0.845
1,000,000	0.0046	stratified	specificity 99%	0.817	0.816	0.811	0.815
1,000,000	0.0092	naive	specificity 99%	0.823	0.823	0.824	0.816
1,000,000	0.0092	stratified	specificity 99%	0.838	0.831	0.830	0.833

Table B.33: Coverage of 95% confidence intervals for specificity at the 99th percentile, logistic regression.

B.6 Negative Predictive Value (NPV)

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 90%	0.972	0.974	0.962	0.970
5,000	0.0046	stratified	NPV 90%	0.980	0.980	0.986	0.986
5,000	0.0092	naive	NPV 90%	0.982	0.980	0.980	0.978
5,000	0.0092	stratified	NPV 90%	0.982	0.988	0.984	0.986
5,000	0.0184	naive	NPV 90%	0.992	0.994	0.990	0.990
5,000	0.0184	stratified	NPV 90%	0.982	0.980	0.988	0.984
10,000	0.0046	naive	NPV 90%	0.978	0.976	0.978	0.976
10,000	0.0046	stratified	NPV 90%	0.990	0.990	0.990	0.988
10,000	0.0092	naive	NPV 90%	0.976	0.978	0.972	0.978
10,000	0.0092	stratified	NPV 90%	0.984	0.986	0.984	0.982
10,000	0.0184	naive	NPV 90%	0.984	0.990	0.978	0.982
10,000	0.0184	stratified	NPV 90%	0.982	0.982	0.984	0.984
50,000	0.0046	naive	NPV 90%	0.990	0.988	0.990	0.990
50,000	0.0046	stratified	NPV 90%	0.988	0.986	0.986	0.984
50,000	0.0092	naive	NPV 90%	0.980	0.980	0.982	0.974
50,000	0.0092	stratified	NPV 90%	0.972	0.970	0.966	0.970
50,000	0.0184	naive	NPV 90%	0.984	0.978	0.976	0.978
50,000	0.0184	stratified	NPV 90%	0.982	0.988	0.980	0.986
100,000	0.0046	naive	NPV 90%	0.980	0.980	0.980	0.982
100,000	0.0046	stratified	NPV 90%	0.978	0.980	0.978	0.978
100,000	0.0092	naive	NPV 90%	0.984	0.988	0.978	0.986
100,000	0.0092	stratified	NPV 90%	0.970	0.970	0.970	0.970
100,000	0.0184	naive	NPV 90%	0.978	0.974	0.976	0.974
100,000	0.0184	stratified	NPV 90%	0.980	0.982	0.984	0.984
1,000,000	0.0046	naive	NPV 90%	0.906	0.910	0.904	0.898
1,000,000	0.0046	stratified	NPV 90%	0.920	0.924	0.918	0.914
1,000,000	0.0092	naive	NPV 90%	0.916	0.914	0.914	0.918
1,000,000	0.0092	stratified	NPV 90%	0.890	0.890	0.884	0.886

Table B.34: Coverage of 95% confidence intervals for NPV at the 90th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 95%	0.982	0.982	0.978	0.974
5,000	0.0046	stratified	NPV 95%	0.992	0.990	0.994	0.992
5,000	0.0092	naive	NPV 95%	0.990	0.990	0.992	0.990
5,000	0.0092	stratified	NPV 95%	0.988	0.988	0.988	0.990
5,000	0.0184	naive	NPV 95%	0.994	0.996	0.994	0.996
5,000	0.0184	stratified	NPV 95%	0.998	0.996	0.998	0.996
10,000	0.0046	naive	NPV 95%	0.990	0.990	0.992	0.992
10,000	0.0046	stratified	NPV 95%	0.998	0.998	0.996	0.996
10,000	0.0092	naive	NPV 95%	0.996	0.996	0.990	0.996
10,000	0.0092	stratified	NPV 95%	0.994	0.992	0.994	0.994
10,000	0.0184	naive	NPV 95%	0.998	0.998	1.000	0.998
10,000	0.0184	stratified	NPV 95%	0.992	0.994	0.994	0.994
50,000	0.0046	naive	NPV 95%	0.996	0.996	0.994	1.000
50,000	0.0046	stratified	NPV 95%	0.990	0.990	0.990	0.992
50,000	0.0092	naive	NPV 95%	0.994	0.994	0.988	0.994
50,000	0.0092	stratified	NPV 95%	0.988	0.990	0.990	0.986
50,000	0.0184	naive	NPV 95%	0.990	0.990	0.992	0.990
50,000	0.0184	stratified	NPV 95%	0.988	0.992	0.990	0.990
100,000	0.0046	naive	NPV 95%	0.988	0.990	0.982	0.982
100,000	0.0046	stratified	NPV 95%	0.988	0.992	0.986	0.988
100,000	0.0092	naive	NPV 95%	0.996	0.994	0.996	0.996
100,000	0.0092	stratified	NPV 95%	0.990	0.990	0.986	0.986
100,000	0.0184	naive	NPV 95%	0.986	0.984	0.986	0.982
100,000	0.0184	stratified	NPV 95%	0.996	0.992	0.988	0.992
1,000,000	0.0046	naive	NPV 95%	0.938	0.942	0.942	0.938
1,000,000	0.0046	stratified	NPV 95%	0.954	0.952	0.948	0.948
1,000,000	0.0092	naive	NPV 95%	0.932	0.928	0.918	0.924
1,000,000	0.0092	stratified	NPV 95%	0.918	0.926	0.920	0.922

Table B.35: Coverage of 95% confidence intervals for NPV at the 95th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 99%	0.996	0.998	0.996	0.996
5,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	0.998
10,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0046	stratified	NPV 99%	1.000	0.998	1.000	1.000
100,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	NPV 99%	0.978	0.982	0.972	0.980
1,000,000	0.0046	stratified	NPV 99%	0.984	0.986	0.984	0.984
1,000,000	0.0092	naive	NPV 99%	0.956	0.956	0.954	0.954
1,000,000	0.0092	stratified	NPV 99%	0.976	0.976	0.970	0.976

Table B.36: Coverage of 95% confidence intervals for NPV at the 99th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 90%	0.966	0.969	0.953	0.960
5,000	0.0046	stratified	NPV 90%	0.978	0.979	0.983	0.983
5,000	0.0092	naive	NPV 90%	0.980	0.977	0.980	0.976
5,000	0.0092	stratified	NPV 90%	0.980	0.978	0.981	0.979
5,000	0.0184	naive	NPV 90%	0.980	0.981	0.976	0.979
5,000	0.0184	stratified	NPV 90%	0.987	0.986	0.984	0.984
10,000	0.0046	naive	NPV 90%	0.984	0.984	0.982	0.987
10,000	0.0046	stratified	NPV 90%	0.980	0.978	0.980	0.978
10,000	0.0092	naive	NPV 90%	0.986	0.980	0.985	0.982
10,000	0.0092	stratified	NPV 90%	0.983	0.983	0.981	0.983
10,000	0.0184	naive	NPV 90%	0.986	0.986	0.988	0.987
10,000	0.0184	stratified	NPV 90%	0.983	0.983	0.981	0.981
50,000	0.0046	naive	NPV 90%	0.975	0.976	0.974	0.975
50,000	0.0046	stratified	NPV 90%	0.984	0.984	0.982	0.987
50,000	0.0092	naive	NPV 90%	0.989	0.988	0.986	0.989
50,000	0.0092	stratified	NPV 90%	0.985	0.983	0.982	0.983
50,000	0.0184	naive	NPV 90%	0.976	0.974	0.975	0.973
50,000	0.0184	stratified	NPV 90%	0.976	0.974	0.976	0.974
100,000	0.0046	naive	NPV 90%	0.973	0.975	0.973	0.974
100,000	0.0046	stratified	NPV 90%	0.969	0.968	0.967	0.969
100,000	0.0092	naive	NPV 90%	0.981	0.983	0.978	0.979
100,000	0.0092	stratified	NPV 90%	0.978	0.981	0.978	0.978
100,000	0.0184	naive	NPV 90%	0.982	0.978	0.981	0.977
100,000	0.0184	stratified	NPV 90%	0.976	0.980	0.972	0.976
1,000,000	0.0046	naive	NPV 90%	0.911	0.913	0.903	0.905
1,000,000	0.0046	stratified	NPV 90%	0.919	0.917	0.913	0.907
1,000,000	0.0092	naive	NPV 90%	0.921	0.918	0.914	0.914
1,000,000	0.0092	stratified	NPV 90%	0.917	0.921	0.908	0.916

Table B.37: Coverage of 95% confidence intervals for NPV at the 90th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 95%	0.980	0.982	0.971	0.977
5,000	0.0046	stratified	NPV 95%	0.987	0.988	0.988	0.989
5,000	0.0092	naive	NPV 95%	0.994	0.992	0.994	0.993
5,000	0.0092	stratified	NPV 95%	0.990	0.992	0.989	0.991
5,000	0.0184	naive	NPV 95%	0.992	0.993	0.992	0.993
5,000	0.0184	stratified	NPV 95%	0.998	0.998	0.996	0.997
10,000	0.0046	naive	NPV 95%	0.994	0.992	0.993	0.993
10,000	0.0046	stratified	NPV 95%	0.991	0.991	0.993	0.990
10,000	0.0092	naive	NPV 95%	0.992	0.996	0.991	0.994
10,000	0.0092	stratified	NPV 95%	0.992	0.992	0.992	0.993
10,000	0.0184	naive	NPV 95%	0.994	0.991	0.993	0.992
10,000	0.0184	stratified	NPV 95%	0.996	0.996	0.995	0.996
50,000	0.0046	naive	NPV 95%	0.992	0.993	0.993	0.994
50,000	0.0046	stratified	NPV 95%	0.990	0.992	0.990	0.992
50,000	0.0092	naive	NPV 95%	0.993	0.993	0.993	0.994
50,000	0.0092	stratified	NPV 95%	0.992	0.992	0.992	0.992
50,000	0.0184	naive	NPV 95%	0.997	0.993	0.996	0.993
50,000	0.0184	stratified	NPV 95%	0.995	0.995	0.995	0.995
100,000	0.0046	naive	NPV 95%	0.991	0.991	0.988	0.989
100,000	0.0046	stratified	NPV 95%	0.993	0.994	0.992	0.989
100,000	0.0092	naive	NPV 95%	0.994	0.994	0.995	0.992
100,000	0.0092	stratified	NPV 95%	0.993	0.995	0.993	0.993
100,000	0.0184	naive	NPV 95%	0.997	0.998	0.996	0.996
100,000	0.0184	stratified	NPV 95%	0.983	0.985	0.984	0.983
1,000,000	0.0046	naive	NPV 95%	0.951	0.952	0.949	0.947
1,000,000	0.0046	stratified	NPV 95%	0.940	0.944	0.932	0.938
1,000,000	0.0092	naive	NPV 95%	0.945	0.945	0.948	0.945
1,000,000	0.0092	stratified	NPV 95%	0.949	0.953	0.950	0.949

Table B.38: Coverage of 95% confidence intervals for NPV at the 95th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 99%	0.996	0.997	0.991	0.991
5,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
5,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0046	naive	NPV 99%	1.000	1.000	0.999	1.000
50,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0092	naive	NPV 99%	1.000	0.999	1.000	1.000
50,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	NPV 99%	0.994	0.994	0.995	0.992
1,000,000	0.0046	stratified	NPV 99%	0.991	0.993	0.991	0.991
1,000,000	0.0092	naive	NPV 99%	0.994	0.996	0.995	0.996
1,000,000	0.0092	stratified	NPV 99%	0.994	0.994	0.990	0.995

Table B.39: Coverage of 95% confidence intervals for NPV at the 99th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 90%	0.991	0.994	0.987	0.991
5,000	0.0046	stratified	NPV 90%	0.998	0.998	0.998	0.998
5,000	0.0092	naive	NPV 90%	0.998	0.998	0.997	0.999
5,000	0.0092	stratified	NPV 90%	0.998	0.995	0.995	0.994
5,000	0.0184	naive	NPV 90%	0.990	0.989	0.985	0.989
5,000	0.0184	stratified	NPV 90%	0.989	0.988	0.986	0.986
10,000	0.0046	naive	NPV 90%	0.997	0.995	0.995	0.995
10,000	0.0046	stratified	NPV 90%	0.999	0.999	0.998	0.998
10,000	0.0092	naive	NPV 90%	0.984	0.987	0.979	0.984
10,000	0.0092	stratified	NPV 90%	0.989	0.988	0.988	0.983
10,000	0.0184	naive	NPV 90%	0.987	0.985	0.980	0.977
10,000	0.0184	stratified	NPV 90%	0.985	0.984	0.980	0.983
50,000	0.0046	naive	NPV 90%	0.981	0.980	0.975	0.979
50,000	0.0046	stratified	NPV 90%	0.972	0.974	0.972	0.970
50,000	0.0092	naive	NPV 90%	0.973	0.974	0.968	0.971
50,000	0.0092	stratified	NPV 90%	0.983	0.986	0.985	0.983
50,000	0.0184	naive	NPV 90%	0.981	0.979	0.979	0.978
50,000	0.0184	stratified	NPV 90%	0.979	0.979	0.979	0.978
100,000	0.0046	naive	NPV 90%	0.974	0.975	0.969	0.975
100,000	0.0046	stratified	NPV 90%	0.978	0.981	0.972	0.976
100,000	0.0092	naive	NPV 90%	0.974	0.972	0.972	0.970
100,000	0.0092	stratified	NPV 90%	0.981	0.974	0.977	0.972
100,000	0.0184	naive	NPV 90%	0.985	0.984	0.983	0.980
100,000	0.0184	stratified	NPV 90%	0.986	0.985	0.985	0.981
1,000,000	0.0046	naive	NPV 90%	0.925	0.918	0.914	0.916
1,000,000	0.0046	stratified	NPV 90%	0.907	0.907	0.899	0.910
1,000,000	0.0092	naive	NPV 90%	0.904	0.901	0.895	0.904
1,000,000	0.0092	stratified	NPV 90%	0.911	0.921	0.907	0.918

Table B.40: Coverage of 95% confidence intervals for NPV at the 90th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 95%	0.994	0.994	0.985	0.992
5,000	0.0046	stratified	NPV 95%	0.998	0.998	0.998	0.998
5,000	0.0092	naive	NPV 95%	0.998	0.998	0.998	0.998
5,000	0.0092	stratified	NPV 95%	0.999	0.999	0.998	0.999
5,000	0.0184	naive	NPV 95%	0.998	0.996	0.997	0.997
5,000	0.0184	stratified	NPV 95%	0.997	0.996	0.995	0.996
10,000	0.0046	naive	NPV 95%	0.998	0.998	0.998	0.997
10,000	0.0046	stratified	NPV 95%	0.999	0.999	0.999	0.999
10,000	0.0092	naive	NPV 95%	0.995	0.994	0.993	0.995
10,000	0.0092	stratified	NPV 95%	0.996	0.996	0.996	0.996
10,000	0.0184	naive	NPV 95%	0.996	0.995	0.996	0.995
10,000	0.0184	stratified	NPV 95%	0.998	0.998	0.997	0.997
50,000	0.0046	naive	NPV 95%	0.996	0.996	0.995	0.995
50,000	0.0046	stratified	NPV 95%	0.993	0.992	0.991	0.990
50,000	0.0092	naive	NPV 95%	0.989	0.990	0.988	0.986
50,000	0.0092	stratified	NPV 95%	0.992	0.992	0.991	0.989
50,000	0.0184	naive	NPV 95%	0.997	0.994	0.996	0.995
50,000	0.0184	stratified	NPV 95%	0.997	0.995	0.993	0.992
100,000	0.0046	naive	NPV 95%	0.993	0.994	0.993	0.992
100,000	0.0046	stratified	NPV 95%	0.988	0.988	0.987	0.987
100,000	0.0092	naive	NPV 95%	0.993	0.994	0.987	0.990
100,000	0.0092	stratified	NPV 95%	0.988	0.988	0.988	0.987
100,000	0.0184	naive	NPV 95%	0.998	0.997	0.996	0.996
100,000	0.0184	stratified	NPV 95%	0.996	0.996	0.996	0.995
1,000,000	0.0046	naive	NPV 95%	0.954	0.949	0.951	0.949
1,000,000	0.0046	stratified	NPV 95%	0.939	0.942	0.938	0.935
1,000,000	0.0092	naive	NPV 95%	0.945	0.947	0.947	0.942
1,000,000	0.0092	stratified	NPV 95%	0.950	0.954	0.948	0.946

Table B.41: Coverage of 95% confidence intervals for NPV at the 95th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	NPV 99%	0.997	0.996	0.996	0.993
5,000	0.0046	stratified	NPV 99%	0.999	0.999	0.999	0.999
5,000	0.0092	naive	NPV 99%	0.999	0.999	0.999	0.999
5,000	0.0092	stratified	NPV 99%	0.999	0.999	0.999	0.999
5,000	0.0184	naive	NPV 99%	1.000	1.000	0.999	1.000
5,000	0.0184	stratified	NPV 99%	0.999	0.999	0.999	0.999
10,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0046	stratified	NPV 99%	0.999	0.999	0.999	0.999
10,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	1.000
10,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0046	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0046	stratified	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
50,000	0.0092	stratified	NPV 99%	1.000	1.000	0.999	1.000
50,000	0.0184	naive	NPV 99%	1.000	1.000	1.000	0.999
50,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0046	naive	NPV 99%	0.999	1.000	0.999	0.999
100,000	0.0046	stratified	NPV 99%	1.000	1.000	0.999	1.000
100,000	0.0092	naive	NPV 99%	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	NPV 99%	1.000	1.000	1.000	0.999
100,000	0.0184	naive	NPV 99%	1.000	1.000	0.999	0.999
100,000	0.0184	stratified	NPV 99%	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	NPV 99%	0.997	0.997	0.994	0.995
1,000,000	0.0046	stratified	NPV 99%	0.995	0.995	0.995	0.992
1,000,000	0.0092	naive	NPV 99%	0.996	0.996	0.995	0.995
1,000,000	0.0092	stratified	NPV 99%	0.994	0.995	0.993	0.996

Table B.42: Coverage of 95% confidence intervals for NPV at the 99th percentile, logistic regression.

B.7 Positive Predictive Value (PPV)

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 90%	0.974	0.980	0.976	0.978
5,000	0.0046	stratified	PPV 90%	0.970	0.970	0.980	0.976
5,000	0.0092	naive	PPV 90%	0.984	0.990	0.986	0.992
5,000	0.0092	stratified	PPV 90%	0.982	0.978	0.986	0.984
5,000	0.0184	naive	PPV 90%	0.992	0.998	0.992	0.996
5,000	0.0184	stratified	PPV 90%	0.994	0.986	0.990	0.986
10,000	0.0046	naive	PPV 90%	0.986	0.986	0.986	0.978
10,000	0.0046	stratified	PPV 90%	0.986	0.990	0.992	0.992
10,000	0.0092	naive	PPV 90%	0.986	0.994	0.990	0.994
10,000	0.0092	stratified	PPV 90%	0.992	0.996	0.994	1.000
10,000	0.0184	naive	PPV 90%	0.998	0.996	0.996	0.996
10,000	0.0184	stratified	PPV 90%	0.998	0.998	0.994	0.994
50,000	0.0046	naive	PPV 90%	0.994	0.996	0.992	0.998
50,000	0.0046	stratified	PPV 90%	0.996	0.994	0.996	0.996
50,000	0.0092	naive	PPV 90%	0.996	0.994	0.996	0.990
50,000	0.0092	stratified	PPV 90%	0.992	0.990	0.990	0.986
50,000	0.0184	naive	PPV 90%	0.990	0.990	0.994	0.992
50,000	0.0184	stratified	PPV 90%	0.992	0.994	0.992	0.992
100,000	0.0046	naive	PPV 90%	0.996	0.996	0.998	0.998
100,000	0.0046	stratified	PPV 90%	0.996	0.998	0.994	0.998
100,000	0.0092	naive	PPV 90%	0.998	0.998	0.998	0.996
100,000	0.0092	stratified	PPV 90%	0.994	0.990	0.994	0.992
100,000	0.0184	naive	PPV 90%	0.990	0.992	0.992	0.990
100,000	0.0184	stratified	PPV 90%	0.988	0.990	0.986	0.990
1,000,000	0.0046	naive	PPV 90%	0.936	0.944	0.940	0.942
1,000,000	0.0046	stratified	PPV 90%	0.940	0.936	0.940	0.934
1,000,000	0.0092	naive	PPV 90%	0.896	0.908	0.898	0.904
1,000,000	0.0092	stratified	PPV 90%	0.870	0.866	0.872	0.868

Table B.43: Coverage of 95% confidence intervals for PPV at the 90th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 95%	0.950	0.962	0.954	0.968
5,000	0.0046	stratified	PPV 95%	0.950	0.950	0.956	0.956
5,000	0.0092	naive	PPV 95%	0.964	0.972	0.966	0.974
5,000	0.0092	stratified	PPV 95%	0.968	0.976	0.976	0.976
5,000	0.0184	naive	PPV 95%	0.976	0.976	0.974	0.978
5,000	0.0184	stratified	PPV 95%	0.982	0.980	0.982	0.982
10,000	0.0046	naive	PPV 95%	0.972	0.982	0.976	0.978
10,000	0.0046	stratified	PPV 95%	0.970	0.976	0.976	0.980
10,000	0.0092	naive	PPV 95%	0.970	0.972	0.972	0.976
10,000	0.0092	stratified	PPV 95%	0.984	0.978	0.982	0.982
10,000	0.0184	naive	PPV 95%	0.992	0.988	0.992	0.986
10,000	0.0184	stratified	PPV 95%	0.990	0.992	0.990	0.994
50,000	0.0046	naive	PPV 95%	0.988	0.996	0.988	0.996
50,000	0.0046	stratified	PPV 95%	0.986	0.988	0.988	0.988
50,000	0.0092	naive	PPV 95%	0.984	0.984	0.984	0.980
50,000	0.0092	stratified	PPV 95%	0.978	0.976	0.980	0.970
50,000	0.0184	naive	PPV 95%	0.982	0.978	0.982	0.974
50,000	0.0184	stratified	PPV 95%	0.986	0.984	0.988	0.982
100,000	0.0046	naive	PPV 95%	0.984	0.986	0.980	0.980
100,000	0.0046	stratified	PPV 95%	0.988	0.986	0.988	0.982
100,000	0.0092	naive	PPV 95%	0.988	0.988	0.986	0.990
100,000	0.0092	stratified	PPV 95%	0.988	0.986	0.986	0.984
100,000	0.0184	naive	PPV 95%	0.970	0.968	0.970	0.968
100,000	0.0184	stratified	PPV 95%	0.968	0.970	0.968	0.970
1,000,000	0.0046	naive	PPV 95%	0.914	0.914	0.906	0.914
1,000,000	0.0046	stratified	PPV 95%	0.934	0.924	0.930	0.932
1,000,000	0.0092	naive	PPV 95%	0.856	0.854	0.856	0.852
1,000,000	0.0092	stratified	PPV 95%	0.894	0.898	0.884	0.890

Table B.44: Coverage of 95% confidence intervals for PPV at the 95th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 99%	0.872	0.878	0.894	0.898
5,000	0.0046	stratified	PPV 99%	0.882	0.876	0.896	0.886
5,000	0.0092	naive	PPV 99%	0.906	0.914	0.912	0.922
5,000	0.0092	stratified	PPV 99%	0.898	0.922	0.906	0.922
5,000	0.0184	naive	PPV 99%	0.930	0.932	0.928	0.940
5,000	0.0184	stratified	PPV 99%	0.930	0.938	0.940	0.938
10,000	0.0046	naive	PPV 99%	0.878	0.932	0.896	0.934
10,000	0.0046	stratified	PPV 99%	0.886	0.910	0.892	0.918
10,000	0.0092	naive	PPV 99%	0.940	0.946	0.936	0.948
10,000	0.0092	stratified	PPV 99%	0.940	0.950	0.944	0.954
10,000	0.0184	naive	PPV 99%	0.936	0.950	0.930	0.956
10,000	0.0184	stratified	PPV 99%	0.946	0.946	0.944	0.946
50,000	0.0046	naive	PPV 99%	0.952	0.956	0.950	0.952
50,000	0.0046	stratified	PPV 99%	0.964	0.974	0.966	0.970
50,000	0.0092	naive	PPV 99%	0.968	0.962	0.966	0.964
50,000	0.0092	stratified	PPV 99%	0.956	0.970	0.950	0.968
50,000	0.0184	naive	PPV 99%	0.964	0.966	0.962	0.964
50,000	0.0184	stratified	PPV 99%	0.964	0.956	0.956	0.946
100,000	0.0046	naive	PPV 99%	0.954	0.948	0.950	0.946
100,000	0.0046	stratified	PPV 99%	0.962	0.962	0.960	0.956
100,000	0.0092	naive	PPV 99%	0.962	0.962	0.958	0.968
100,000	0.0092	stratified	PPV 99%	0.950	0.956	0.954	0.948
100,000	0.0184	naive	PPV 99%	0.948	0.944	0.940	0.942
100,000	0.0184	stratified	PPV 99%	0.952	0.958	0.944	0.950
1,000,000	0.0046	naive	PPV 99%	0.768	0.764	0.762	0.758
1,000,000	0.0046	stratified	PPV 99%	0.808	0.812	0.806	0.812
1,000,000	0.0092	naive	PPV 99%	0.718	0.716	0.716	0.704
1,000,000	0.0092	stratified	PPV 99%	0.736	0.736	0.736	0.726

Table B.45: Coverage of 95% confidence intervals for PPV at the 99th percentile, random forest.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 90%	0.957	0.966	0.962	0.941
5,000	0.0046	stratified	PPV 90%	0.976	0.986	0.976	0.991
5,000	0.0092	naive	PPV 90%	0.988	0.989	0.991	0.985
5,000	0.0092	stratified	PPV 90%	0.989	0.990	0.988	0.990
5,000	0.0184	naive	PPV 90%	0.984	0.988	0.985	0.987
5,000	0.0184	stratified	PPV 90%	0.990	0.992	0.990	0.991
10,000	0.0046	naive	PPV 90%	0.980	0.995	0.986	0.992
10,000	0.0046	stratified	PPV 90%	0.989	0.992	0.989	0.990
10,000	0.0092	naive	PPV 90%	0.994	0.996	0.994	0.995
10,000	0.0092	stratified	PPV 90%	0.991	0.993	0.990	0.993
10,000	0.0184	naive	PPV 90%	0.992	0.993	0.992	0.991
10,000	0.0184	stratified	PPV 90%	0.989	0.990	0.989	0.988
50,000	0.0046	naive	PPV 90%	0.991	0.995	0.991	0.992
50,000	0.0046	stratified	PPV 90%	0.993	0.993	0.990	0.991
50,000	0.0092	naive	PPV 90%	0.999	0.998	0.999	0.997
50,000	0.0092	stratified	PPV 90%	0.997	0.998	0.997	0.998
50,000	0.0184	naive	PPV 90%	0.991	0.990	0.989	0.986
50,000	0.0184	stratified	PPV 90%	0.995	0.992	0.992	0.990
100,000	0.0046	naive	PPV 90%	0.996	0.997	0.995	0.996
100,000	0.0046	stratified	PPV 90%	0.993	0.992	0.992	0.991
100,000	0.0092	naive	PPV 90%	0.997	0.997	0.995	0.996
100,000	0.0092	stratified	PPV 90%	0.997	0.996	0.998	0.994
100,000	0.0184	naive	PPV 90%	0.997	0.997	0.996	0.996
100,000	0.0184	stratified	PPV 90%	0.992	0.992	0.990	0.991
1,000,000	0.0046	naive	PPV 90%	0.970	0.972	0.971	0.973
1,000,000	0.0046	stratified	PPV 90%	0.970	0.973	0.969	0.970
1,000,000	0.0092	naive	PPV 90%	0.970	0.970	0.967	0.971
1,000,000	0.0092	stratified	PPV 90%	0.969	0.968	0.967	0.970

Table B.46: Coverage of 95% confidence intervals for PPV at the 90th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 95%	0.930	0.948	0.939	0.937
5,000	0.0046	stratified	PPV 95%	0.945	0.968	0.953	0.976
5,000	0.0092	naive	PPV 95%	0.968	0.986	0.975	0.981
5,000	0.0092	stratified	PPV 95%	0.971	0.978	0.971	0.979
5,000	0.0184	naive	PPV 95%	0.968	0.972	0.973	0.978
5,000	0.0184	stratified	PPV 95%	0.977	0.984	0.979	0.983
10,000	0.0046	naive	PPV 95%	0.961	0.987	0.969	0.982
10,000	0.0046	stratified	PPV 95%	0.978	0.987	0.982	0.988
10,000	0.0092	naive	PPV 95%	0.976	0.984	0.977	0.985
10,000	0.0092	stratified	PPV 95%	0.979	0.982	0.982	0.986
10,000	0.0184	naive	PPV 95%	0.972	0.971	0.972	0.971
10,000	0.0184	stratified	PPV 95%	0.981	0.984	0.981	0.982
50,000	0.0046	naive	PPV 95%	0.983	0.983	0.980	0.981
50,000	0.0046	stratified	PPV 95%	0.985	0.984	0.985	0.983
50,000	0.0092	naive	PPV 95%	0.989	0.987	0.989	0.986
50,000	0.0092	stratified	PPV 95%	0.983	0.984	0.984	0.985
50,000	0.0184	naive	PPV 95%	0.975	0.974	0.975	0.974
50,000	0.0184	stratified	PPV 95%	0.984	0.983	0.982	0.981
100,000	0.0046	naive	PPV 95%	0.988	0.991	0.986	0.990
100,000	0.0046	stratified	PPV 95%	0.987	0.987	0.985	0.988
100,000	0.0092	naive	PPV 95%	0.988	0.988	0.986	0.986
100,000	0.0092	stratified	PPV 95%	0.987	0.986	0.986	0.985
100,000	0.0184	naive	PPV 95%	0.985	0.982	0.980	0.986
100,000	0.0184	stratified	PPV 95%	0.975	0.976	0.975	0.974
1,000,000	0.0046	naive	PPV 95%	0.949	0.949	0.946	0.949
1,000,000	0.0046	stratified	PPV 95%	0.932	0.935	0.933	0.935
1,000,000	0.0092	naive	PPV 95%	0.946	0.952	0.939	0.947
1,000,000	0.0092	stratified	PPV 95%	0.944	0.944	0.942	0.940

Table B.47: Coverage of 95% confidence intervals for PPV at the 95th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 99%	0.845	0.870	0.863	0.884
5,000	0.0046	stratified	PPV 99%	0.841	0.886	0.870	0.898
5,000	0.0092	naive	PPV 99%	0.916	0.925	0.922	0.938
5,000	0.0092	stratified	PPV 99%	0.921	0.929	0.929	0.933
5,000	0.0184	naive	PPV 99%	0.942	0.934	0.944	0.938
5,000	0.0184	stratified	PPV 99%	0.949	0.946	0.953	0.948
10,000	0.0046	naive	PPV 99%	0.922	0.962	0.930	0.966
10,000	0.0046	stratified	PPV 99%	0.922	0.932	0.920	0.934
10,000	0.0092	naive	PPV 99%	0.948	0.947	0.946	0.945
10,000	0.0092	stratified	PPV 99%	0.933	0.941	0.936	0.938
10,000	0.0184	naive	PPV 99%	0.932	0.948	0.936	0.951
10,000	0.0184	stratified	PPV 99%	0.952	0.960	0.950	0.961
50,000	0.0046	naive	PPV 99%	0.966	0.972	0.965	0.970
50,000	0.0046	stratified	PPV 99%	0.959	0.958	0.958	0.957
50,000	0.0092	naive	PPV 99%	0.960	0.964	0.956	0.966
50,000	0.0092	stratified	PPV 99%	0.952	0.950	0.952	0.948
50,000	0.0184	naive	PPV 99%	0.953	0.959	0.948	0.956
50,000	0.0184	stratified	PPV 99%	0.950	0.960	0.946	0.955
100,000	0.0046	naive	PPV 99%	0.963	0.968	0.964	0.962
100,000	0.0046	stratified	PPV 99%	0.965	0.973	0.955	0.968
100,000	0.0092	naive	PPV 99%	0.953	0.950	0.949	0.955
100,000	0.0092	stratified	PPV 99%	0.961	0.962	0.962	0.960
100,000	0.0184	naive	PPV 99%	0.943	0.949	0.942	0.949
100,000	0.0184	stratified	PPV 99%	0.942	0.938	0.940	0.939
1,000,000	0.0046	naive	PPV 99%	0.865	0.864	0.861	0.862
1,000,000	0.0046	stratified	PPV 99%	0.885	0.886	0.886	0.883
1,000,000	0.0092	naive	PPV 99%	0.896	0.896	0.885	0.894
1,000,000	0.0092	stratified	PPV 99%	0.887	0.890	0.884	0.887

Table B.48: Coverage of 95% confidence intervals for PPV at the 99th percentile, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 90%	0.946	0.932	0.941	0.928
5,000	0.0046	stratified	PPV 90%	0.954	0.921	0.948	0.930
5,000	0.0092	naive	PPV 90%	0.908	0.914	0.912	0.917
5,000	0.0092	stratified	PPV 90%	0.915	0.919	0.910	0.923
5,000	0.0184	naive	PPV 90%	0.932	0.929	0.932	0.930
5,000	0.0184	stratified	PPV 90%	0.925	0.936	0.926	0.932
10,000	0.0046	naive	PPV 90%	0.893	0.898	0.903	0.904
10,000	0.0046	stratified	PPV 90%	0.921	0.922	0.924	0.923
10,000	0.0092	naive	PPV 90%	0.930	0.933	0.930	0.934
10,000	0.0092	stratified	PPV 90%	0.950	0.949	0.954	0.954
10,000	0.0184	naive	PPV 90%	0.974	0.978	0.977	0.979
10,000	0.0184	stratified	PPV 90%	0.964	0.966	0.963	0.967
50,000	0.0046	naive	PPV 90%	0.983	0.985	0.982	0.982
50,000	0.0046	stratified	PPV 90%	0.980	0.981	0.981	0.982
50,000	0.0092	naive	PPV 90%	0.986	0.987	0.983	0.985
50,000	0.0092	stratified	PPV 90%	0.996	0.997	0.995	0.996
50,000	0.0184	naive	PPV 90%	0.994	0.993	0.994	0.992
50,000	0.0184	stratified	PPV 90%	0.990	0.991	0.992	0.992
100,000	0.0046	naive	PPV 90%	0.990	0.990	0.989	0.991
100,000	0.0046	stratified	PPV 90%	0.994	0.995	0.994	0.994
100,000	0.0092	naive	PPV 90%	0.990	0.991	0.989	0.989
100,000	0.0092	stratified	PPV 90%	0.992	0.992	0.993	0.993
100,000	0.0184	naive	PPV 90%	0.995	0.994	0.996	0.991
100,000	0.0184	stratified	PPV 90%	0.996	0.995	0.992	0.995
1,000,000	0.0046	naive	PPV 90%	0.976	0.976	0.973	0.972
1,000,000	0.0046	stratified	PPV 90%	0.969	0.966	0.962	0.965
1,000,000	0.0092	naive	PPV 90%	0.974	0.972	0.965	0.971
1,000,000	0.0092	stratified	PPV 90%	0.973	0.973	0.973	0.970

Table B.49: Coverage of 95% confidence intervals for PPV at the 90th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 95%	0.931	0.927	0.925	0.927
5,000	0.0046	stratified	PPV 95%	0.924	0.909	0.920	0.913
5,000	0.0092	naive	PPV 95%	0.862	0.880	0.872	0.894
5,000	0.0092	stratified	PPV 95%	0.885	0.900	0.893	0.904
5,000	0.0184	naive	PPV 95%	0.906	0.906	0.912	0.914
5,000	0.0184	stratified	PPV 95%	0.896	0.910	0.900	0.914
10,000	0.0046	naive	PPV 95%	0.882	0.883	0.891	0.902
10,000	0.0046	stratified	PPV 95%	0.899	0.899	0.910	0.907
10,000	0.0092	naive	PPV 95%	0.904	0.917	0.912	0.921
10,000	0.0092	stratified	PPV 95%	0.916	0.913	0.928	0.918
10,000	0.0184	naive	PPV 95%	0.944	0.948	0.946	0.948
10,000	0.0184	stratified	PPV 95%	0.946	0.951	0.949	0.953
50,000	0.0046	naive	PPV 95%	0.974	0.974	0.975	0.972
50,000	0.0046	stratified	PPV 95%	0.966	0.965	0.968	0.963
50,000	0.0092	naive	PPV 95%	0.974	0.978	0.972	0.977
50,000	0.0092	stratified	PPV 95%	0.979	0.982	0.976	0.979
50,000	0.0184	naive	PPV 95%	0.983	0.982	0.983	0.980
50,000	0.0184	stratified	PPV 95%	0.968	0.969	0.969	0.970
100,000	0.0046	naive	PPV 95%	0.980	0.978	0.976	0.978
100,000	0.0046	stratified	PPV 95%	0.981	0.977	0.978	0.978
100,000	0.0092	naive	PPV 95%	0.975	0.977	0.976	0.978
100,000	0.0092	stratified	PPV 95%	0.977	0.976	0.975	0.974
100,000	0.0184	naive	PPV 95%	0.980	0.979	0.978	0.976
100,000	0.0184	stratified	PPV 95%	0.985	0.980	0.983	0.982
1,000,000	0.0046	naive	PPV 95%	0.955	0.949	0.947	0.947
1,000,000	0.0046	stratified	PPV 95%	0.938	0.943	0.935	0.936
1,000,000	0.0092	naive	PPV 95%	0.939	0.935	0.940	0.938
1,000,000	0.0092	stratified	PPV 95%	0.937	0.936	0.929	0.931

Table B.50: Coverage of 95% confidence intervals for PPV at the 95th percentile, logistic regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	PPV 99%	0.860	0.870	0.864	0.871
5,000	0.0046	stratified	PPV 99%	0.858	0.859	0.857	0.855
5,000	0.0092	naive	PPV 99%	0.807	0.843	0.827	0.858
5,000	0.0092	stratified	PPV 99%	0.830	0.857	0.844	0.882
5,000	0.0184	naive	PPV 99%	0.884	0.903	0.889	0.918
5,000	0.0184	stratified	PPV 99%	0.881	0.899	0.893	0.913
10,000	0.0046	naive	PPV 99%	0.843	0.832	0.871	0.858
10,000	0.0046	stratified	PPV 99%	0.860	0.829	0.881	0.853
10,000	0.0092	naive	PPV 99%	0.899	0.904	0.913	0.925
10,000	0.0092	stratified	PPV 99%	0.887	0.893	0.900	0.906
10,000	0.0184	naive	PPV 99%	0.907	0.920	0.911	0.928
10,000	0.0184	stratified	PPV 99%	0.906	0.918	0.911	0.923
50,000	0.0046	naive	PPV 99%	0.938	0.949	0.944	0.952
50,000	0.0046	stratified	PPV 99%	0.942	0.937	0.942	0.939
50,000	0.0092	naive	PPV 99%	0.938	0.943	0.938	0.944
50,000	0.0092	stratified	PPV 99%	0.942	0.947	0.944	0.946
50,000	0.0184	naive	PPV 99%	0.953	0.953	0.944	0.955
50,000	0.0184	stratified	PPV 99%	0.944	0.943	0.938	0.943
100,000	0.0046	naive	PPV 99%	0.947	0.941	0.947	0.946
100,000	0.0046	stratified	PPV 99%	0.948	0.947	0.952	0.940
100,000	0.0092	naive	PPV 99%	0.951	0.948	0.946	0.942
100,000	0.0092	stratified	PPV 99%	0.947	0.950	0.947	0.948
100,000	0.0184	naive	PPV 99%	0.962	0.961	0.956	0.953
100,000	0.0184	stratified	PPV 99%	0.946	0.950	0.945	0.946
1,000,000	0.0046	naive	PPV 99%	0.870	0.870	0.866	0.868
1,000,000	0.0046	stratified	PPV 99%	0.873	0.881	0.873	0.876
1,000,000	0.0092	naive	PPV 99%	0.880	0.883	0.884	0.880
1,000,000	0.0092	stratified	PPV 99%	0.879	0.880	0.876	0.871

Table B.51: Coverage of 95% confidence intervals for PPV at the 99th percentile, logistic regression.

Appendix C

BIAS, MSE, AND CV-VALUE VARIABILITY

C.1 Area under the ROC curve (AUC)

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
AUC	Random Forest	5,000	0.0046	naive	-0.0041 (0.0027)	0.0039 (0.00026)	0.0039	0.086	0.595	0.96
AUC	Random Forest	5,000	0.0046	stratified	-0.0014 (0.0024)	0.0032 (0.00021)	0.0032	0.077	0.621	0.953
AUC	Random Forest	5,000	0.0092	naive	-0.0016 (0.0018)	0.0017 (0.00014)	0.0017	0.054	0.659	0.931
AUC	Random Forest	5,000	0.0092	stratified	-0.003 (0.0015)	0.0012 (7.1e-05)	0.0012	0.05	0.74	0.923
AUC	Random Forest	5,000	0.0184	naive	0.00081 (0.0011)	0.00057 (3.7e-05)	0.00057	0.03	0.764	0.902
AUC	Random Forest	5,000	0.0184	stratified	-0.00054 (0.001)	0.00054 (3.3e-05)	0.00054	0.032	0.772	0.907
AUC	Random Forest	10,000	0.0046	naive	0.0024 (0.0017)	0.0014 (8.2e-05)	0.0014	0.052	0.71	0.92
AUC	Random Forest	10,000	0.0046	stratified	-0.0016 (0.0015)	0.0011 (7.2e-05)	0.0011	0.044	0.707	0.937
AUC	Random Forest	10,000	0.0092	naive	-0.0011 (0.0012)	0.0007 (4.4e-05)	0.0007	0.037	0.753	0.925
AUC	Random Forest	10,000	0.0092	stratified	-0.0007 (0.0011)	0.00063 (4.1e-05)	0.00063	0.034	0.747	0.903
AUC	Random Forest	10,000	0.0184	naive	-0.0014 (0.00075)	0.00028 (1.8e-05)	0.00028	0.024	0.798	0.899
AUC	Random Forest	10,000	0.0184	stratified	0.00065 (0.00071)	0.00025 (1.6e-05)	0.00025	0.021	0.802	0.9
AUC	Random Forest	50,000	0.0046	naive	-0.00052 (0.00066)	0.00021 (1.3e-05)	0.00021	0.02	0.803	0.891
AUC	Random Forest	50,000	0.0046	stratified	-0.00036 (0.00061)	0.00018 (1.1e-05)	0.00018	0.019	0.806	0.888
AUC	Random Forest	50,000	0.0092	naive	-0.0011 (0.00044)	9.2e-05 (5.7e-06)	9.1e-05	0.013	0.829	0.887
AUC	Random Forest	50,000	0.0092	stratified	-0.00057 (0.00045)	9e-05 (6.1e-06)	9e-05	0.013	0.822	0.887
AUC	Random Forest	50,000	0.0184	naive	-0.00087 (0.0003)	4.1e-05 (2.8e-06)	4e-05	0.0082	0.839	0.882
AUC	Random Forest	50,000	0.0184	stratified	-0.001 (0.0003)	3.9e-05 (2.3e-06)	3.8e-05	0.0084	0.844	0.879
AUC	Random Forest	100,000	0.0046	naive	-0.0011 (0.00045)	9e-05 (6.8e-06)	8.9e-05	0.013	0.814	0.883
AUC	Random Forest	100,000	0.0046	stratified	-0.00073 (0.00043)	8.5e-05 (5.6e-06)	8.4e-05	0.012	0.827	0.883
AUC	Random Forest	100,000	0.0092	naive	-0.00033 (0.0003)	4e-05 (2.7e-06)	4e-05	0.0079	0.841	0.88
AUC	Random Forest	100,000	0.0092	stratified	-0.00041 (0.00032)	4.2e-05 (2.6e-06)	4.2e-05	0.0091	0.843	0.882
AUC	Random Forest	100,000	0.0184	naive	-0.00082 (0.00022)	2.1e-05 (1.4e-06)	2e-05	0.0059	0.849	0.878
AUC	Random Forest	100,000	0.0184	stratified	-0.00095 (0.00022)	2.1e-05 (1.3e-06)	2e-05	0.0061	0.853	0.879
AUC	Random Forest	1,000,000	0.0046	naive	-0.00021 (0.00018)	6.3e-06 (4.1e-07)	6.3e-06	0.0034	0.859	0.874
AUC	Random Forest	1,000,000	0.0046	stratified	-0.00046 (0.00017)	6.1e-06 (4.3e-07)	5.9e-06	0.0032	0.857	0.874
AUC	Random Forest	1,000,000	0.0092	naive	-0.00034 (0.00012)	2.4e-06 (1.7e-07)	2.3e-06	0.0018	0.862	0.872
AUC	Random Forest	1,000,000	0.0092	stratified	-0.00053 (0.00013)	2.8e-06 (1.7e-07)	2.5e-06	0.0022	0.863	0.872

Table C.1: Bias, MSE, and Variability for AUC, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
AUC	Ridge	5,000	0.0046	naive	-0.0012 (0.0017)	0.0031 (0.00015)	0.0031	0.076	0.555	0.956
AUC	Ridge	5,000	0.0046	stratified	-0.0019 (0.0016)	0.0025 (0.00012)	0.0025	0.066	0.612	0.969
AUC	Ridge	5,000	0.0092	naive	0.0022 (0.0012)	0.0014 (6e-05)	0.0014	0.053	0.682	0.918
AUC	Ridge	5,000	0.0092	stratified	-0.00074 (0.0011)	0.0011 (5.3e-05)	0.0011	0.045	0.713	0.931
AUC	Ridge	5,000	0.0184	naive	-0.00026 (0.00078)	0.00059 (2.6e-05)	0.00059	0.033	0.737	0.916
AUC	Ridge	5,000	0.0184	stratified	-0.00088 (0.00073)	0.00052 (2.4e-05)	0.00052	0.029	0.754	0.911
AUC	Ridge	10,000	0.0046	naive	0.00099 (0.0013)	0.0015 (7e-05)	0.0015	0.054	0.68	0.927
AUC	Ridge	10,000	0.0046	stratified	-0.0011 (0.0011)	0.0011 (5.2e-05)	0.0011	0.044	0.705	0.932
AUC	Ridge	10,000	0.0092	naive	-0.0018 (0.00075)	0.00054 (2.3e-05)	0.00054	0.033	0.76	0.901
AUC	Ridge	10,000	0.0092	stratified	-0.001 (0.00075)	0.00055 (2.2e-05)	0.00055	0.032	0.766	0.903
AUC	Ridge	10,000	0.0184	naive	-0.0014 (0.00053)	0.00028 (1.2e-05)	0.00028	0.023	0.792	0.89
AUC	Ridge	10,000	0.0184	stratified	-0.00065 (0.00051)	0.00025 (1.1e-05)	0.00025	0.021	0.8	0.894
AUC	Ridge	50,000	0.0046	naive	-0.0011 (0.00049)	0.00022 (9.2e-06)	0.00022	0.02	0.802	0.89
AUC	Ridge	50,000	0.0046	stratified	-0.0012 (0.00044)	0.00018 (8.8e-06)	0.00018	0.019	0.785	0.887
AUC	Ridge	50,000	0.0092	naive	-0.00068 (0.00033)	9.7e-05 (4.5e-06)	9.6e-05	0.013	0.822	0.883
AUC	Ridge	50,000	0.0092	stratified	-0.001 (0.00032)	9.3e-05 (4.1e-06)	9.2e-05	0.013	0.824	0.883
AUC	Ridge	50,000	0.0184	naive	-0.00059 (0.00022)	4.2e-05 (1.9e-06)	4.2e-05	0.0088	0.837	0.881
AUC	Ridge	50,000	0.0184	stratified	-0.00065 (0.00022)	4.3e-05 (1.9e-06)	4.2e-05	0.0091	0.839	0.883
AUC	Ridge	100,000	0.0046	naive	-0.00036 (0.00033)	9.3e-05 (4.1e-06)	9.3e-05	0.013	0.823	0.885
AUC	Ridge	100,000	0.0046	stratified	-0.00031 (0.00034)	0.0001 (4.8e-06)	0.0001	0.013	0.819	0.885
AUC	Ridge	100,000	0.0092	naive	-0.00084 (0.00022)	4.2e-05 (1.9e-06)	4.1e-05	0.0085	0.839	0.881
AUC	Ridge	100,000	0.0092	stratified	-0.00035 (0.00022)	4.1e-05 (1.9e-06)	4.1e-05	0.0085	0.836	0.879
AUC	Ridge	100,000	0.0184	naive	-0.00043 (0.00016)	2e-05 (9.1e-07)	2e-05	0.0058	0.848	0.877
AUC	Ridge	100,000	0.0184	stratified	-0.00023 (0.00016)	2e-05 (9.2e-07)	2e-05	0.0058	0.85	0.877
AUC	Ridge	1,000,000	0.0046	naive	0.00022 (0.00014)	7e-06 (3e-07)	7e-06	0.0037	0.855	0.873
AUC	Ridge	1,000,000	0.0046	stratified	-0.00023 (0.00013)	7.6e-06 (3.1e-07)	7.5e-06	0.0039	0.856	0.871
AUC	Ridge	1,000,000	0.0092	naive	-0.00026 (9.1e-05)	2.6e-06 (1.1e-07)	2.5e-06	0.0022	0.86	0.869
AUC	Ridge	1,000,000	0.0092	stratified	-0.00019 (9e-05)	2.5e-06 (1.1e-07)	2.5e-06	0.0022	0.86	0.87

Table C.2: Bias, MSE, and Variability for AUC, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
AUC	Logistic	5,000	0.0046	naive	0.014 (0.0018)	0.0045 (0.00022)	0.0043	0.074	0.272	0.724
AUC	Logistic	5,000	0.0046	stratified	0.017 (0.0017)	0.004 (0.0002)	0.0038	0.065	0.275	0.711
AUC	Logistic	5,000	0.0092	naive	-0.016 (0.0023)	0.0045 (0.0002)	0.0042	0.085	0.333	0.765
AUC	Logistic	5,000	0.0092	stratified	-0.014 (0.0023)	0.0041 (0.0002)	0.0039	0.081	0.309	0.722
AUC	Logistic	5,000	0.0184	naive	-0.019 (0.0017)	0.0023 (0.00011)	0.0019	0.059	0.48	0.781
AUC	Logistic	5,000	0.0184	stratified	-0.022 (0.0017)	0.0023 (1e-04)	0.0018	0.057	0.516	0.776
AUC	Logistic	10,000	0.0046	naive	-0.02 (0.0026)	0.0052 (0.00023)	0.0048	0.089	0.305	0.737
AUC	Logistic	10,000	0.0046	stratified	-0.017 (0.0023)	0.0043 (0.00019)	0.004	0.082	0.349	0.702
AUC	Logistic	10,000	0.0092	naive	-0.022 (0.0016)	0.0021 (9.7e-05)	0.0017	0.053	0.531	0.771
AUC	Logistic	10,000	0.0092	stratified	-0.021 (0.0016)	0.002 (0.0001)	0.0016	0.051	0.479	0.764
AUC	Logistic	10,000	0.0184	naive	-0.012 (0.00096)	0.00071 (3.3e-05)	0.00056	0.033	0.682	0.834
AUC	Logistic	10,000	0.0184	stratified	-0.014 (0.00094)	0.00073 (3.3e-05)	0.00053	0.032	0.674	0.833
AUC	Logistic	50,000	0.0046	naive	-0.0096 (0.00078)	0.00047 (2.1e-05)	0.00038	0.026	0.723	0.855
AUC	Logistic	50,000	0.0046	stratified	-0.0092 (0.00075)	0.00042 (1.8e-05)	0.00034	0.023	0.736	0.853
AUC	Logistic	50,000	0.0092	naive	-0.0041 (0.00044)	0.00016 (7.8e-06)	0.00014	0.015	0.789	0.869
AUC	Logistic	50,000	0.0092	stratified	-0.0035 (0.00041)	0.00014 (5.9e-06)	0.00013	0.015	0.802	0.868
AUC	Logistic	50,000	0.0184	naive	-0.0017 (0.00025)	5.4e-05 (2.5e-06)	5.1e-05	0.0097	0.83	0.874
AUC	Logistic	50,000	0.0184	stratified	-0.0017 (0.00024)	5e-05 (2.3e-06)	4.7e-05	0.0094	0.83	0.877
AUC	Logistic	100,000	0.0046	naive	-0.0036 (0.00042)	0.00014 (6.9e-06)	0.00013	0.015	0.796	0.875
AUC	Logistic	100,000	0.0046	stratified	-0.0033 (0.00042)	0.00013 (6e-06)	0.00012	0.015	0.801	0.869
AUC	Logistic	100,000	0.0092	naive	-0.0014 (0.00025)	5.1e-05 (2.2e-06)	4.9e-05	0.0092	0.835	0.875
AUC	Logistic	100,000	0.0092	stratified	-0.0017 (0.00025)	5.3e-05 (2.4e-06)	5e-05	0.0094	0.829	0.875
AUC	Logistic	100,000	0.0184	naive	-0.00063 (0.00016)	2e-05 (8.5e-07)	1.9e-05	0.0063	0.849	0.874
AUC	Logistic	100,000	0.0184	stratified	-0.00082 (0.00015)	1.9e-05 (9e-07)	1.9e-05	0.0056	0.848	0.876
AUC	Logistic	1,000,000	0.0046	naive	-0.0002 (0.00013)	6.7e-06 (3.1e-07)	6.7e-06	0.0035	0.855	0.875
AUC	Logistic	1,000,000	0.0046	stratified	-9.3e-05 (0.00013)	6.6e-06 (2.9e-07)	6.6e-06	0.0034	0.858	0.874
AUC	Logistic	1,000,000	0.0092	naive	5.7e-05 (9.3e-05)	2.7e-06 (1.3e-07)	2.7e-06	0.0022	0.861	0.872
AUC	Logistic	1,000,000	0.0092	stratified	-5.1e-05 (9.2e-05)	2.8e-06 (1.2e-07)	2.8e-06	0.0023	0.861	0.871

Table C.3: Bias, MSE, and Variability for AUC, logistic regression.

C.2 Brier Score

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
Brier	Random Forest	5,000	0.0046	naive	0.00051 (2.2e-05)	5.1e-07 (4.2e-08)	2.5e-07	0.00055	0.00408	0.00745
Brier	Random Forest	5,000	0.0046	stratified	1.3e-05 (3.6e-06)	7.8e-09 (8.4e-10)	7.6e-09	0.0001	0.00392	0.00461
Brier	Random Forest	5,000	0.0092	naive	0.00011 (1.3e-05)	1e-07 (1.2e-08)	9.2e-08	0.00027	0.00761	0.00998
Brier	Random Forest	5,000	0.0092	stratified	2.7e-05 (8.1e-06)	4.3e-08 (3e-09)	4.2e-08	0.00028	0.00772	0.00906
Brier	Random Forest	5,000	0.0184	naive	9.1e-05 (1.6e-05)	1.8e-07 (1.2e-08)	1.7e-07	0.00056	0.0146	0.018
Brier	Random Forest	5,000	0.0184	stratified	6.7e-05 (1.5e-05)	1.6e-07 (8.9e-09)	1.5e-07	0.00055	0.0153	0.0173
Brier	Random Forest	10,000	0.0046	naive	5.9e-05 (7e-06)	2.9e-08 (4.1e-09)	2.5e-08	6e-05	0.00421	0.00555
Brier	Random Forest	10,000	0.0046	stratified	8.9e-06 (1.8e-06)	2.4e-09 (2e-10)	2.3e-09	5.5e-05	0.00426	0.00456
Brier	Random Forest	10,000	0.0092	naive	4.4e-05 (6e-06)	2.5e-08 (2.2e-09)	2.4e-08	0.00021	0.00804	0.00942
Brier	Random Forest	10,000	0.0092	stratified	1.6e-05 (5.6e-06)	2e-08 (1.2e-09)	2e-08	0.00019	0.0081	0.00891
Brier	Random Forest	10,000	0.0184	naive	9.5e-05 (1.1e-05)	8.9e-08 (5.6e-09)	8e-08	0.00037	0.0151	0.017
Brier	Random Forest	10,000	0.0184	stratified	5.3e-05 (1.2e-05)	9.1e-08 (5.6e-09)	8.8e-08	0.00041	0.0152	0.0169
Brier	Random Forest	50,000	0.0046	naive	1.2e-05 (9.5e-07)	8.5e-10 (4.6e-11)	7.2e-10	3.8e-05	0.00435	0.0045
Brier	Random Forest	50,000	0.0046	stratified	1.1e-05 (8.9e-07)	7.6e-10 (4.6e-11)	6.4e-10	3.2e-05	0.00435	0.0045
Brier	Random Forest	50,000	0.0092	naive	2.7e-05 (2.1e-06)	3.9e-09 (2.3e-10)	3.2e-09	7.6e-05	0.00838	0.00875
Brier	Random Forest	50,000	0.0092	stratified	2.8e-05 (2.1e-06)	4.1e-09 (2.5e-10)	3.4e-09	7.5e-05	0.00841	0.00874
Brier	Random Forest	50,000	0.0184	naive	5.1e-05 (4.4e-06)	1.7e-08 (1.1e-09)	1.4e-08	0.00015	0.0158	0.0167
Brier	Random Forest	50,000	0.0184	stratified	4.7e-05 (4.5e-06)	1.7e-08 (1.1e-09)	1.5e-08	0.00016	0.0157	0.0167
Brier	Random Forest	100,000	0.0046	naive	1.2e-05 (7.3e-07)	5.3e-10 (2.9e-11)	3.9e-10	2.8e-05	0.00437	0.00447
Brier	Random Forest	100,000	0.0046	stratified	1.1e-05 (7.2e-07)	5e-10 (2.9e-11)	3.8e-10	2.6e-05	0.00435	0.00447
Brier	Random Forest	100,000	0.0092	naive	1.6e-05 (1.6e-06)	2e-09 (1.2e-10)	1.7e-09	5.5e-05	0.0084	0.00866
Brier	Random Forest	100,000	0.0092	stratified	1.3e-05 (1.7e-06)	2.1e-09 (1.3e-10)	1.9e-09	5.2e-05	0.00834	0.00864
Brier	Random Forest	100,000	0.0184	naive	3.5e-05 (3.5e-06)	8.9e-09 (5.4e-10)	7.6e-09	0.00011	0.0158	0.0164
Brier	Random Forest	100,000	0.0184	stratified	3.9e-05 (3.7e-06)	9.6e-09 (5.4e-10)	8e-09	0.00013	0.0158	0.0163
Brier	Random Forest	1,000,000	0.0046	naive	2.4e-06 (3.2e-07)	3.8e-11 (2.5e-12)	3.2e-11	6.8e-06	0.00439	0.00443
Brier	Random Forest	1,000,000	0.0046	stratified	2.2e-06 (3e-07)	3.2e-11 (1.9e-12)	2.7e-11	6.9e-06	0.0044	0.00443
Brier	Random Forest	1,000,000	0.0092	naive	6.9e-06 (7.7e-07)	1.8e-10 (1e-11)	1.4e-10	1.6e-05	0.0085	0.00857
Brier	Random Forest	1,000,000	0.0092	stratified	6.8e-06 (7.3e-07)	1.7e-10 (9.4e-12)	1.2e-10	1.5e-05	0.0085	0.00856

Table C.4: Bias, MSE, and Variability for Brier score, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
Brier	Ridge	5,000	0.0046	naive	0.00049 (1.5e-05)	4.5e-07 (2.3e-08)	2.1e-07	0.00051	0.00426	0.00758
Brier	Ridge	5,000	0.0046	stratified	1.1e-05 (1.2e-06)	2.1e-09 (2.6e-10)	2e-09	3.1e-05	0.00416	0.00462
Brier	Ridge	5,000	0.0092	naive	0.00012 (1e-05)	1.3e-07 (1e-08)	1.2e-07	0.00031	0.00756	0.0107
Brier	Ridge	5,000	0.0092	stratified	3.1e-05 (5.2e-06)	4e-08 (2e-09)	3.9e-08	0.00026	0.0079	0.00917
Brier	Ridge	5,000	0.0184	naive	5.7e-05 (1.6e-05)	2.9e-07 (1.3e-08)	2.9e-07	0.00071	0.0141	0.0176
Brier	Ridge	5,000	0.0184	stratified	6e-05 (1.6e-05)	2.8e-07 (1.2e-08)	2.8e-07	0.00071	0.014	0.0178
Brier	Ridge	10,000	0.0046	naive	4.8e-05 (4.4e-06)	2.2e-08 (2.2e-09)	2e-08	8e-05	0.00421	0.00507
Brier	Ridge	10,000	0.0046	stratified	1.5e-05 (2.4e-06)	7e-09 (3.8e-10)	6.8e-09	0.00011	0.00405	0.00469
Brier	Ridge	10,000	0.0092	naive	4.4e-05 (6.6e-06)	4.7e-08 (2e-09)	4.5e-08	0.00028	0.00775	0.00922
Brier	Ridge	10,000	0.0092	stratified	2.9e-05 (7.4e-06)	5.5e-08 (2.5e-09)	5.4e-08	0.0003	0.00768	0.00916
Brier	Ridge	10,000	0.0184	naive	5.7e-05 (1.3e-05)	1.7e-07 (7.7e-09)	1.6e-07	0.00053	0.0147	0.0171
Brier	Ridge	10,000	0.0184	stratified	5e-05 (1.3e-05)	1.7e-07 (7.1e-09)	1.7e-07	0.00057	0.0149	0.0171
Brier	Ridge	50,000	0.0046	naive	1.7e-05 (2e-06)	3.5e-09 (1.4e-10)	3.2e-09	7.9e-05	0.00415	0.00453
Brier	Ridge	50,000	0.0046	stratified	1.5e-05 (2e-06)	3.3e-09 (1.4e-10)	3.1e-09	7.2e-05	0.00417	0.00453
Brier	Ridge	50,000	0.0092	naive	2.7e-05 (3.6e-06)	1.2e-08 (5.3e-10)	1.1e-08	0.00014	0.00804	0.00875
Brier	Ridge	50,000	0.0092	stratified	2.5e-05 (3.7e-06)	1.2e-08 (5.1e-10)	1.1e-08	0.00013	0.00811	0.00876
Brier	Ridge	50,000	0.0184	naive	3.4e-05 (6.1e-06)	3.2e-08 (1.4e-09)	3.1e-08	0.00024	0.0153	0.0164
Brier	Ridge	50,000	0.0184	stratified	4e-05 (6.2e-06)	3.4e-08 (1.5e-09)	3.2e-08	0.00024	0.0153	0.0165
Brier	Ridge	100,000	0.0046	naive	1.3e-05 (1.4e-06)	1.7e-09 (7.3e-11)	1.5e-09	5e-05	0.00424	0.00447
Brier	Ridge	100,000	0.0046	stratified	9.5e-06 (1.5e-06)	1.8e-09 (7.6e-11)	1.7e-09	5.9e-05	0.00421	0.00449
Brier	Ridge	100,000	0.0092	naive	2.1e-05 (2.6e-06)	5.8e-09 (2.5e-10)	5.4e-09	9.9e-05	0.00813	0.0086
Brier	Ridge	100,000	0.0092	stratified	1.5e-05 (2.7e-06)	5.7e-09 (2.6e-10)	5.5e-09	0.0001	0.00813	0.00862
Brier	Ridge	100,000	0.0184	naive	2e-05 (4.5e-06)	1.6e-08 (7.1e-10)	1.6e-08	0.00017	0.0153	0.0161
Brier	Ridge	100,000	0.0184	stratified	1.5e-05 (4.5e-06)	1.6e-08 (7.1e-10)	1.6e-08	0.00018	0.0153	0.0161
Brier	Ridge	1,000,000	0.0046	naive	7.4e-07 (6.1e-07)	1.5e-10 (6.8e-12)	1.5e-10	1.7e-05	0.00428	0.00438
Brier	Ridge	1,000,000	0.0046	stratified	-5.1e-07 (6e-07)	1.4e-10 (6.3e-12)	1.4e-10	1.5e-05	0.00429	0.00437
Brier	Ridge	1,000,000	0.0092	naive	1.3e-06 (1.1e-06)	3.6e-10 (1.6e-11)	3.6e-10	2.6e-05	0.00824	0.00837
Brier	Ridge	1,000,000	0.0092	stratified	1.9e-06 (1.1e-06)	3.6e-10 (1.6e-11)	3.6e-10	2.6e-05	0.00824	0.00836

Table C.5: Bias, MSE, and Variability for Brier score, ridge regression.

				Coverage			
Training set size	Rate	Sampling strategy	Metric	Wald, across	Wald, within	Percentile, across	Percentile, within
5,000	0.0046	naive	Brier	0.403	0.396	0.392	0.391
5,000	0.0046	stratified	Brier	0.432	0.430	0.422	0.423
5,000	0.0092	naive	Brier	0.816	0.815	0.798	0.803
5,000	0.0092	stratified	Brier	0.837	0.842	0.828	0.831
5,000	0.0184	naive	Brier	0.971	0.970	0.969	0.966
5,000	0.0184	stratified	Brier	0.951	0.948	0.946	0.944
10,000	0.0046	naive	Brier	0.969	0.970	0.963	0.965
10,000	0.0046	stratified	Brier	0.959	0.953	0.953	0.952
10,000	0.0092	naive	Brier	0.993	0.993	0.993	0.993
10,000	0.0092	stratified	Brier	0.989	0.989	0.989	0.989
10,000	0.0184	naive	Brier	0.999	0.999	0.999	0.999
10,000	0.0184	stratified	Brier	0.999	0.999	0.997	0.998
50,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
50,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	naive	Brier	1.000	1.000	1.000	1.000
100,000	0.0184	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0046	stratified	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	naive	Brier	1.000	1.000	1.000	1.000
1,000,000	0.0092	stratified	Brier	1.000	1.000	1.000	1.000

Table C.6: Bias, MSE, and Variability for Brier score, logistic regression.

C.3 Accuracy

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 90%	Random Forest	5,000	0.0046	naive	-0.00044 (0.00022)	9.5e-06 (6.8e-07)	9.3e-06	0.0038	0.885	0.906
accuracy 90%	Random Forest	5,000	0.0046	stratified	8e-06 (0.00022)	6.5e-06 (4.2e-07)	6.5e-06	0.0034	0.887	0.902
accuracy 90%	Random Forest	5,000	0.0092	naive	-0.00016 (0.00021)	6.2e-06 (4e-07)	6.2e-06	0.0032	0.887	0.903
accuracy 90%	Random Forest	5,000	0.0092	stratified	-0.00027 (0.00021)	6e-06 (3.9e-07)	6e-06	0.0032	0.889	0.904
accuracy 90%	Random Forest	5,000	0.0184	naive	-0.00015 (0.00021)	6.1e-06 (4e-07)	6.1e-06	0.0032	0.892	0.907
accuracy 90%	Random Forest	5,000	0.0184	stratified	0.00011 (0.0002)	5.4e-06 (3.5e-07)	5.4e-06	0.0032	0.892	0.906
accuracy 90%	Random Forest	10,000	0.0046	naive	-0.00031 (0.00014)	2.1e-06 (1.3e-07)	2e-06	0.0019	0.895	0.904
accuracy 90%	Random Forest	10,000	0.0046	stratified	-0.00019 (0.00015)	1.9e-06 (1.3e-07)	1.9e-06	0.0018	0.896	0.904
accuracy 90%	Random Forest	10,000	0.0092	naive	-5.6e-05 (0.00014)	2.9e-06 (1.9e-07)	2.9e-06	0.0023	0.89	0.901
accuracy 90%	Random Forest	10,000	0.0092	stratified	-0.00013 (0.00015)	2.6e-06 (1.5e-07)	2.5e-06	0.0022	0.892	0.901
accuracy 90%	Random Forest	10,000	0.0184	naive	-0.00018 (0.00014)	3.1e-06 (2.1e-07)	3.1e-06	0.0024	0.894	0.905
accuracy 90%	Random Forest	10,000	0.0184	stratified	0.00015 (0.00014)	2.9e-06 (2e-07)	2.9e-06	0.0022	0.894	0.905
accuracy 90%	Random Forest	50,000	0.0046	naive	-5.6e-05 (6.5e-05)	2.9e-07 (1.9e-08)	2.9e-07	0.00074	0.898	0.902
accuracy 90%	Random Forest	50,000	0.0046	stratified	-6.5e-05 (6.4e-05)	2.8e-07 (1.7e-08)	2.7e-07	0.00068	0.899	0.902
accuracy 90%	Random Forest	50,000	0.0092	naive	1.5e-05 (6.6e-05)	3.1e-07 (1.9e-08)	3.1e-07	0.00074	0.9	0.903
accuracy 90%	Random Forest	50,000	0.0092	stratified	-2e-06 (6.4e-05)	3e-07 (2e-08)	3e-07	0.00075	0.9	0.904
accuracy 90%	Random Forest	50,000	0.0184	naive	-3.5e-05 (6.2e-05)	4.2e-07 (2.7e-08)	4.2e-07	0.0009	0.901	0.905
accuracy 90%	Random Forest	50,000	0.0184	stratified	-0.00012 (6e-05)	4.5e-07 (2.6e-08)	4.4e-07	0.00088	0.902	0.905
accuracy 90%	Random Forest	100,000	0.0046	naive	9.9e-06 (4.9e-05)	1.7e-07 (1.1e-08)	1.7e-07	0.00056	0.899	0.902
accuracy 90%	Random Forest	100,000	0.0046	stratified	-5.9e-05 (4.5e-05)	1.8e-07 (1e-08)	1.7e-07	0.00056	0.899	0.902
accuracy 90%	Random Forest	100,000	0.0092	naive	5.2e-05 (4.7e-05)	1.8e-07 (1.2e-08)	1.8e-07	0.00054	0.9	0.903
accuracy 90%	Random Forest	100,000	0.0092	stratified	-1.8e-05 (4.6e-05)	1.7e-07 (9.9e-09)	1.7e-07	0.00055	0.901	0.903
accuracy 90%	Random Forest	100,000	0.0184	naive	-6.8e-05 (4.3e-05)	2.2e-07 (1.5e-08)	2.1e-07	0.00061	0.902	0.905
accuracy 90%	Random Forest	100,000	0.0184	stratified	-5.4e-05 (4.6e-05)	2.1e-07 (1.4e-08)	2.1e-07	0.00061	0.903	0.905
accuracy 90%	Random Forest	1,000,000	0.0046	naive	-1.3e-05 (2e-05)	1.3e-08 (1e-09)	1.3e-08	0.00014	0.901	0.902
accuracy 90%	Random Forest	1,000,000	0.0046	stratified	-3.2e-05 (2e-05)	1.2e-08 (7.7e-10)	1.1e-08	0.00014	0.901	0.901
accuracy 90%	Random Forest	1,000,000	0.0092	naive	-4e-05 (1.9e-05)	1.5e-08 (9.8e-10)	1.4e-08	0.00015	0.902	0.903
accuracy 90%	Random Forest	1,000,000	0.0092	stratified	-2.6e-05 (2e-05)	1.3e-08 (8.3e-10)	1.2e-08	0.00014	0.902	0.903

Table C.7: Bias, MSE, and Variability accuracy at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 95%	Random Forest	5,000	0.0046	naive	-0.00056 (0.00017)	5.7e-06 (4e-07)	5.4e-06	0.003	0.937	0.954
accuracy 95%	Random Forest	5,000	0.0046	stratified	-0.00024 (0.00016)	3.5e-06 (2.3e-07)	3.4e-06	0.0026	0.939	0.951
accuracy 95%	Random Forest	5,000	0.0092	naive	-0.00014 (0.00015)	3.4e-06 (2.6e-07)	3.4e-06	0.0026	0.939	0.952
accuracy 95%	Random Forest	5,000	0.0092	stratified	-0.00038 (0.00015)	3.5e-06 (2.1e-07)	3.3e-06	0.0024	0.942	0.951
accuracy 95%	Random Forest	5,000	0.0184	naive	-8.7e-05 (0.00016)	4.1e-06 (2.6e-07)	4.1e-06	0.0027	0.941	0.953
accuracy 95%	Random Forest	5,000	0.0184	stratified	-6.7e-06 (0.00014)	4.1e-06 (2.3e-07)	4.1e-06	0.0028	0.941	0.952
accuracy 95%	Random Forest	10,000	0.0046	naive	-8.8e-06 (0.00011)	1.3e-06 (8.7e-08)	1.3e-06	0.0015	0.945	0.953
accuracy 95%	Random Forest	10,000	0.0046	stratified	-0.00015 (0.00011)	1.4e-06 (9.4e-08)	1.4e-06	0.0016	0.945	0.953
accuracy 95%	Random Forest	10,000	0.0092	naive	-7.5e-05 (0.0001)	1.7e-06 (1.1e-07)	1.7e-06	0.0017	0.944	0.952
accuracy 95%	Random Forest	10,000	0.0092	stratified	2.5e-06 (9.8e-05)	1.6e-06 (9.7e-08)	1.6e-06	0.0017	0.944	0.951
accuracy 95%	Random Forest	10,000	0.0184	naive	-0.0001 (9.4e-05)	1.8e-06 (1.1e-07)	1.8e-06	0.0017	0.944	0.951
accuracy 95%	Random Forest	10,000	0.0184	stratified	-0.00014 (9.9e-05)	1.8e-06 (1.1e-07)	1.8e-06	0.0018	0.944	0.952
accuracy 95%	Random Forest	50,000	0.0046	naive	7.1e-06 (4.8e-05)	2e-07 (1.2e-08)	2e-07	0.0006	0.948	0.951
accuracy 95%	Random Forest	50,000	0.0046	stratified	-3.7e-05 (4.5e-05)	1.9e-07 (1.3e-08)	1.9e-07	0.00058	0.948	0.951
accuracy 95%	Random Forest	50,000	0.0092	naive	-9.9e-05 (4.6e-05)	2.5e-07 (1.6e-08)	2.4e-07	0.0007	0.948	0.951
accuracy 95%	Random Forest	50,000	0.0092	stratified	-6.5e-05 (4.6e-05)	2.4e-07 (1.4e-08)	2.3e-07	0.0007	0.948	0.951
accuracy 95%	Random Forest	50,000	0.0184	naive	-9.8e-05 (4.6e-05)	3.4e-07 (2.1e-08)	3.4e-07	0.00076	0.947	0.951
accuracy 95%	Random Forest	50,000	0.0184	stratified	-7.2e-05 (4.5e-05)	3.8e-07 (2.3e-08)	3.8e-07	0.00086	0.947	0.95
accuracy 95%	Random Forest	100,000	0.0046	naive	-2.3e-05 (3.4e-05)	1e-07 (6.3e-09)	1e-07	0.00039	0.949	0.95
accuracy 95%	Random Forest	100,000	0.0046	stratified	-8.4e-05 (3.2e-05)	1.3e-07 (7.5e-09)	1.2e-07	0.00045	0.949	0.951
accuracy 95%	Random Forest	100,000	0.0092	naive	1.7e-05 (3.3e-05)	1.3e-07 (8.2e-09)	1.3e-07	0.00048	0.948	0.95
accuracy 95%	Random Forest	100,000	0.0092	stratified	-1.1e-05 (3.3e-05)	1.2e-07 (7.2e-09)	1.2e-07	0.00045	0.949	0.951
accuracy 95%	Random Forest	100,000	0.0184	naive	-5.7e-05 (3.3e-05)	2.1e-07 (1.3e-08)	2e-07	0.00061	0.947	0.951
accuracy 95%	Random Forest	100,000	0.0184	stratified	-4.1e-06 (3.2e-05)	2e-07 (1.2e-08)	2e-07	0.00061	0.948	0.95
accuracy 95%	Random Forest	1,000,000	0.0046	naive	-1.7e-05 (1.4e-05)	9.2e-09 (5.4e-10)	8.9e-09	0.00014	0.95	0.95
accuracy 95%	Random Forest	1,000,000	0.0046	stratified	-1.2e-05 (1.4e-05)	8.2e-09 (5.3e-10)	8.1e-09	0.00012	0.95	0.95
accuracy 95%	Random Forest	1,000,000	0.0092	naive	-2e-05 (1.4e-05)	1e-08 (6.4e-10)	9.9e-09	0.00014	0.949	0.95
accuracy 95%	Random Forest	1,000,000	0.0092	stratified	-4.8e-05 (1.4e-05)	1.2e-08 (7.8e-10)	9.6e-09	0.00013	0.949	0.95

Table C.8: Bias, MSE, and Variability accuracy at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 99%	Random Forest	5,000	0.0046	naive	-0.00063 (8.3e-05)	2.1e-06 (1.9e-07)	1.7e-06	0.0017	0.98	0.99
accuracy 99%	Random Forest	5,000	0.0046	stratified	3.4e-05 (6.8e-05)	9.9e-07 (6.1e-08)	9.9e-07	0.0014	0.985	0.99
accuracy 99%	Random Forest	5,000	0.0092	naive	-0.00013 (6.6e-05)	1.3e-06 (7.6e-08)	1.3e-06	0.0016	0.982	0.988
accuracy 99%	Random Forest	5,000	0.0092	stratified	-0.00011 (6.6e-05)	1.4e-06 (9.1e-08)	1.4e-06	0.0016	0.981	0.989
accuracy 99%	Random Forest	5,000	0.0184	naive	-0.00011 (6e-05)	1.6e-06 (9.8e-08)	1.6e-06	0.0018	0.976	0.983
accuracy 99%	Random Forest	5,000	0.0184	stratified	-5.8e-05 (5.7e-05)	1.4e-06 (8.4e-08)	1.4e-06	0.0016	0.976	0.983
accuracy 99%	Random Forest	10,000	0.0046	naive	-5e-05 (5.3e-05)	5.3e-07 (3.7e-08)	5.3e-07	0.00097	0.984	0.989
accuracy 99%	Random Forest	10,000	0.0046	stratified	-0.00014 (4.9e-05)	5e-07 (3.3e-08)	4.8e-07	0.001	0.985	0.989
accuracy 99%	Random Forest	10,000	0.0092	naive	-0.00014 (4.7e-05)	6.4e-07 (4.1e-08)	6.2e-07	0.0011	0.983	0.988
accuracy 99%	Random Forest	10,000	0.0092	stratified	-2.2e-05 (4.6e-05)	5.6e-07 (3.5e-08)	5.6e-07	0.001	0.983	0.988
accuracy 99%	Random Forest	10,000	0.0184	naive	-0.00012 (4.2e-05)	6.6e-07 (4.4e-08)	6.5e-07	0.0011	0.977	0.982
accuracy 99%	Random Forest	10,000	0.0184	stratified	-4.7e-05 (4.3e-05)	7.3e-07 (4.5e-08)	7.3e-07	0.0012	0.977	0.982
accuracy 99%	Random Forest	50,000	0.0046	naive	-1.9e-05 (2.1e-05)	8.2e-08 (4.7e-09)	8.1e-08	0.0004	0.987	0.988
accuracy 99%	Random Forest	50,000	0.0046	stratified	-2.7e-05 (2.1e-05)	7.1e-08 (4.8e-09)	7.1e-08	0.00036	0.987	0.988
accuracy 99%	Random Forest	50,000	0.0092	naive	-3.1e-05 (2e-05)	9.6e-08 (6e-09)	9.5e-08	0.0004	0.984	0.986
accuracy 99%	Random Forest	50,000	0.0092	stratified	-4.2e-05 (1.9e-05)	1e-07 (6.6e-09)	1e-07	0.00042	0.984	0.986
accuracy 99%	Random Forest	50,000	0.0184	naive	-3.2e-05 (1.9e-05)	1.4e-07 (8.8e-09)	1.4e-07	0.0005	0.978	0.98
accuracy 99%	Random Forest	50,000	0.0184	stratified	-3.9e-05 (2e-05)	1.4e-07 (9.6e-09)	1.4e-07	0.00045	0.978	0.981
accuracy 99%	Random Forest	100,000	0.0046	naive	-3.1e-05 (1.5e-05)	4.3e-08 (2.5e-09)	4.2e-08	0.00029	0.987	0.988
accuracy 99%	Random Forest	100,000	0.0046	stratified	-4.3e-05 (1.5e-05)	4.1e-08 (2.9e-09)	3.9e-08	0.00024	0.987	0.988
accuracy 99%	Random Forest	100,000	0.0092	naive	-1.1e-05 (1.4e-05)	5.2e-08 (3.3e-09)	5.2e-08	0.00031	0.984	0.986
accuracy 99%	Random Forest	100,000	0.0092	stratified	-1.2e-05 (1.5e-05)	5e-08 (3.4e-09)	5e-08	0.00028	0.984	0.986
accuracy 99%	Random Forest	100,000	0.0184	naive	-3.1e-05 (1.5e-05)	6.8e-08 (4.5e-09)	6.7e-08	0.00035	0.979	0.98
accuracy 99%	Random Forest	100,000	0.0184	stratified	-3.1e-05 (1.4e-05)	7.2e-08 (4.5e-09)	7.1e-08	0.00036	0.978	0.98
accuracy 99%	Random Forest	1,000,000	0.0046	naive	-2.6e-06 (6.8e-06)	3.3e-09 (2.2e-10)	3.3e-09	7.4e-05	0.987	0.988
accuracy 99%	Random Forest	1,000,000	0.0046	stratified	3.7e-06 (6.3e-06)	2.6e-09 (1.7e-10)	2.6e-09	6.7e-05	0.987	0.988
accuracy 99%	Random Forest	1,000,000	0.0092	naive	-7.3e-06 (6.4e-06)	3.9e-09 (2.2e-10)	3.8e-09	8.7e-05	0.985	0.985
accuracy 99%	Random Forest	1,000,000	0.0092	stratified	-8.8e-06 (6.4e-06)	4.1e-09 (2.6e-10)	4e-09	8.6e-05	0.985	0.985

Table C.9: Bias, MSE, and Variability accuracy at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 90%	Ridge	5,000	0.0046	naive	-0.00024 (0.00015)	4.9e-06 (2.7e-07)	4.9e-06	0.0028	0.889	0.907
accuracy 90%	Ridge	5,000	0.0046	stratified	-2.8e-05 (0.00014)	2.7e-06 (1.3e-07)	2.7e-06	0.0022	0.893	0.905
accuracy 90%	Ridge	5,000	0.0092	naive	-0.00015 (0.00014)	3.7e-06 (1.7e-07)	3.7e-06	0.0026	0.894	0.907
accuracy 90%	Ridge	5,000	0.0092	stratified	-0.00023 (0.00014)	3.3e-06 (1.5e-07)	3.3e-06	0.0024	0.895	0.906
accuracy 90%	Ridge	5,000	0.0184	naive	-0.00017 (0.00013)	5e-06 (2.3e-07)	5e-06	0.0028	0.894	0.909
accuracy 90%	Ridge	5,000	0.0184	stratified	-0.00017 (0.00014)	4.7e-06 (2e-07)	4.7e-06	0.003	0.896	0.909
accuracy 90%	Ridge	10,000	0.0046	naive	-2.4e-05 (9.9e-05)	1.3e-06 (5.7e-08)	1.3e-06	0.0015	0.896	0.904
accuracy 90%	Ridge	10,000	0.0046	stratified	-0.00016 (0.0001)	1.4e-06 (6.1e-08)	1.4e-06	0.0016	0.897	0.904
accuracy 90%	Ridge	10,000	0.0092	naive	-4.5e-05 (9.7e-05)	1.6e-06 (7.2e-08)	1.6e-06	0.0016	0.897	0.905
accuracy 90%	Ridge	10,000	0.0092	stratified	-6.7e-05 (9.7e-05)	1.8e-06 (7.8e-08)	1.8e-06	0.0019	0.897	0.906
accuracy 90%	Ridge	10,000	0.0184	naive	-1.3e-05 (9.6e-05)	2.5e-06 (1.1e-07)	2.5e-06	0.0021	0.899	0.908
accuracy 90%	Ridge	10,000	0.0184	stratified	-0.00022 (9.9e-05)	2.4e-06 (1.1e-07)	2.4e-06	0.002	0.898	0.907
accuracy 90%	Ridge	50,000	0.0046	naive	-4.8e-05 (4.6e-05)	2.6e-07 (1.2e-08)	2.5e-07	0.00068	0.899	0.903
accuracy 90%	Ridge	50,000	0.0046	stratified	-3.4e-05 (4.4e-05)	2.7e-07 (1.2e-08)	2.7e-07	0.0007	0.899	0.902
accuracy 90%	Ridge	50,000	0.0092	naive	-4.4e-05 (4.5e-05)	3.3e-07 (1.4e-08)	3.2e-07	0.0008	0.901	0.904
accuracy 90%	Ridge	50,000	0.0092	stratified	-2.2e-05 (4.5e-05)	3.3e-07 (1.5e-08)	3.3e-07	0.0008	0.9	0.904
accuracy 90%	Ridge	50,000	0.0184	naive	-5.6e-05 (4.5e-05)	4.6e-07 (2.1e-08)	4.6e-07	0.00092	0.902	0.907
accuracy 90%	Ridge	50,000	0.0184	stratified	-4e-05 (4.4e-05)	4.7e-07 (2.1e-08)	4.6e-07	0.0009	0.902	0.907
accuracy 90%	Ridge	100,000	0.0046	naive	-3.5e-06 (3.3e-05)	1.2e-07 (5.1e-09)	1.2e-07	0.00047	0.9	0.902
accuracy 90%	Ridge	100,000	0.0046	stratified	3.4e-05 (3.2e-05)	1.2e-07 (5.4e-09)	1.2e-07	0.00048	0.9	0.902
accuracy 90%	Ridge	100,000	0.0092	naive	2.6e-05 (3.2e-05)	1.5e-07 (7.3e-09)	1.5e-07	0.00052	0.901	0.904
accuracy 90%	Ridge	100,000	0.0092	stratified	-7.4e-06 (3.3e-05)	1.5e-07 (6.7e-09)	1.5e-07	0.00053	0.901	0.904
accuracy 90%	Ridge	100,000	0.0184	naive	1.8e-05 (3.2e-05)	1.9e-07 (9e-09)	1.9e-07	0.00061	0.903	0.906
accuracy 90%	Ridge	100,000	0.0184	stratified	3.8e-05 (3.2e-05)	2e-07 (9.4e-09)	2e-07	0.00059	0.903	0.906
accuracy 90%	Ridge	1,000,000	0.0046	naive	-6e-06 (1.4e-05)	6.6e-09 (2.8e-10)	6.5e-09	0.00011	0.901	0.902
accuracy 90%	Ridge	1,000,000	0.0046	stratified	1.2e-05 (1.4e-05)	6.9e-09 (3e-10)	6.7e-09	0.00011	0.901	0.902
accuracy 90%	Ridge	1,000,000	0.0092	naive	-3.6e-05 (1.4e-05)	8.8e-09 (3.9e-10)	7.6e-09	0.00011	0.902	0.903
accuracy 90%	Ridge	1,000,000	0.0092	stratified	-3.8e-05 (1.3e-05)	9.3e-09 (4e-10)	7.9e-09	0.00012	0.902	0.903

Table C.10: Bias, MSE, and Variability for accuracy at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 95%	Ridge	5,000	0.0046	naive	-0.00044 (0.00011)	3.9e-06 (1.9e-07)	3.7e-06	0.0026	0.942	0.956
accuracy 95%	Ridge	5,000	0.0046	stratified	-0.00011 (0.0001)	2e-06 (9e-08)	2e-06	0.002	0.944	0.953
accuracy 95%	Ridge	5,000	0.0092	naive	-0.00017 (9.9e-05)	2.8e-06 (1.3e-07)	2.8e-06	0.0022	0.943	0.954
accuracy 95%	Ridge	5,000	0.0092	stratified	-0.00013 (0.0001)	2.7e-06 (1.2e-07)	2.7e-06	0.0022	0.943	0.953
accuracy 95%	Ridge	5,000	0.0184	naive	-7.8e-05 (9.8e-05)	4.3e-06 (1.9e-07)	4.3e-06	0.0028	0.94	0.955
accuracy 95%	Ridge	5,000	0.0184	stratified	2.8e-06 (0.0001)	4.1e-06 (1.8e-07)	4.1e-06	0.0028	0.941	0.954
accuracy 95%	Ridge	10,000	0.0046	naive	-2.3e-05 (7.4e-05)	9.9e-07 (4.2e-08)	9.9e-07	0.0014	0.946	0.952
accuracy 95%	Ridge	10,000	0.0046	stratified	-1.5e-05 (6.9e-05)	8.4e-07 (3.7e-08)	8.4e-07	0.0012	0.946	0.952
accuracy 95%	Ridge	10,000	0.0092	naive	5.9e-06 (7.2e-05)	1.2e-06 (5.9e-08)	1.2e-06	0.0014	0.945	0.953
accuracy 95%	Ridge	10,000	0.0092	stratified	-0.00011 (7.3e-05)	1.4e-06 (5.9e-08)	1.4e-06	0.0016	0.945	0.952
accuracy 95%	Ridge	10,000	0.0184	naive	3.5e-05 (7.2e-05)	2e-06 (8.5e-08)	2e-06	0.002	0.944	0.953
accuracy 95%	Ridge	10,000	0.0184	stratified	-0.00015 (7e-05)	2e-06 (9e-08)	2e-06	0.0019	0.944	0.952
accuracy 95%	Ridge	50,000	0.0046	naive	-3.2e-05 (3.3e-05)	1.9e-07 (7.9e-09)	1.8e-07	0.0006	0.949	0.951
accuracy 95%	Ridge	50,000	0.0046	stratified	-6.5e-06 (3.3e-05)	1.8e-07 (7.9e-09)	1.8e-07	0.00056	0.949	0.951
accuracy 95%	Ridge	50,000	0.0092	naive	-4.1e-05 (3.2e-05)	2.4e-07 (1e-08)	2.4e-07	0.00064	0.948	0.951
accuracy 95%	Ridge	50,000	0.0092	stratified	-4.4e-05 (3.1e-05)	2.4e-07 (1.2e-08)	2.4e-07	0.0006	0.948	0.951
accuracy 95%	Ridge	50,000	0.0184	naive	-5.9e-05 (3.2e-05)	3.9e-07 (1.9e-08)	3.8e-07	0.00078	0.947	0.952
accuracy 95%	Ridge	50,000	0.0184	stratified	-4e-05 (3.2e-05)	3.7e-07 (1.7e-08)	3.7e-07	0.00078	0.947	0.951
accuracy 95%	Ridge	100,000	0.0046	naive	-6.9e-06 (2.3e-05)	8.8e-08 (3.8e-09)	8.8e-08	0.00041	0.949	0.951
accuracy 95%	Ridge	100,000	0.0046	stratified	8.5e-06 (2.3e-05)	8.5e-08 (3.8e-09)	8.5e-08	0.00039	0.949	0.951
accuracy 95%	Ridge	100,000	0.0092	naive	5.3e-06 (2.3e-05)	1.2e-07 (5.2e-09)	1.2e-07	0.00046	0.949	0.951
accuracy 95%	Ridge	100,000	0.0092	stratified	-2.6e-05 (2.4e-05)	1.2e-07 (5.2e-09)	1.2e-07	0.00046	0.949	0.951
accuracy 95%	Ridge	100,000	0.0184	naive	3.7e-06 (2.3e-05)	1.9e-07 (8.1e-09)	1.9e-07	0.00058	0.948	0.951
accuracy 95%	Ridge	100,000	0.0184	stratified	5.1e-07 (2.3e-05)	1.9e-07 (8.9e-09)	1.9e-07	0.00057	0.948	0.951
accuracy 95%	Ridge	1,000,000	0.0046	naive	-6.6e-06 (9.9e-06)	5.3e-09 (2.4e-10)	5.3e-09	9.7e-05	0.95	0.95
accuracy 95%	Ridge	1,000,000	0.0046	stratified	-1.5e-05 (9.6e-06)	5.9e-09 (2.6e-10)	5.6e-09	0.0001	0.95	0.95
accuracy 95%	Ridge	1,000,000	0.0092	naive	-8.6e-06 (9.5e-06)	7e-09 (3.1e-10)	6.9e-09	0.00012	0.95	0.95
accuracy 95%	Ridge	1,000,000	0.0092	stratified	-1.3e-05 (9.3e-06)	6.7e-09 (3e-10)	6.5e-09	0.00011	0.95	0.95

Table C.11: Bias, MSE, and Variability for accuracy at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 99%	Ridge	5,000	0.0046	naive	-0.00043 (5.1e-05)	1.5e-06 (7.4e-08)	1.3e-06	0.0014	0.982	0.99
accuracy 99%	Ridge	5,000	0.0046	stratified	-0.00011 (4.7e-05)	9.1e-07 (3.9e-08)	9e-07	0.0014	0.984	0.99
accuracy 99%	Ridge	5,000	0.0092	naive	-0.00016 (4.6e-05)	1.4e-06 (5.9e-08)	1.3e-06	0.0016	0.981	0.988
accuracy 99%	Ridge	5,000	0.0092	stratified	-2.8e-05 (4.6e-05)	1.3e-06 (5.8e-08)	1.3e-06	0.0016	0.981	0.988
accuracy 99%	Ridge	5,000	0.0184	naive	-2.7e-05 (4.3e-05)	1.8e-06 (8.1e-08)	1.8e-06	0.0018	0.975	0.983
accuracy 99%	Ridge	5,000	0.0184	stratified	-8.1e-05 (4.5e-05)	1.9e-06 (8.2e-08)	1.8e-06	0.0018	0.974	0.983
accuracy 99%	Ridge	10,000	0.0046	naive	-2.2e-05 (3.3e-05)	4.2e-07 (1.9e-08)	4.2e-07	0.0008	0.985	0.989
accuracy 99%	Ridge	10,000	0.0046	stratified	1.4e-05 (3.4e-05)	4.5e-07 (2.1e-08)	4.5e-07	0.0009	0.985	0.989
accuracy 99%	Ridge	10,000	0.0092	naive	-1.4e-05 (3.3e-05)	6.9e-07 (3e-08)	6.9e-07	0.0011	0.982	0.988
accuracy 99%	Ridge	10,000	0.0092	stratified	-9e-05 (3.4e-05)	7.1e-07 (3.3e-08)	7.1e-07	0.0011	0.982	0.988
accuracy 99%	Ridge	10,000	0.0184	naive	2e-05 (3.3e-05)	1.1e-06 (4.9e-08)	1.1e-06	0.0013	0.976	0.983
accuracy 99%	Ridge	10,000	0.0184	stratified	-7.5e-05 (3.2e-05)	9.9e-07 (4e-08)	9.8e-07	0.0014	0.976	0.982
accuracy 99%	Ridge	50,000	0.0046	naive	-1.9e-05 (1.4e-05)	7.3e-08 (3.4e-09)	7.3e-08	0.00036	0.987	0.988
accuracy 99%	Ridge	50,000	0.0046	stratified	-1.8e-05 (1.5e-05)	8.2e-08 (3.4e-09)	8.1e-08	0.0004	0.987	0.988
accuracy 99%	Ridge	50,000	0.0092	naive	-3.6e-05 (1.5e-05)	1.3e-07 (5.6e-09)	1.2e-07	0.00048	0.984	0.986
accuracy 99%	Ridge	50,000	0.0092	stratified	-2.7e-05 (1.5e-05)	1.3e-07 (5.7e-09)	1.3e-07	0.00048	0.984	0.986
accuracy 99%	Ridge	50,000	0.0184	naive	-1.4e-05 (1.4e-05)	1.8e-07 (8.1e-09)	1.8e-07	0.00058	0.978	0.981
accuracy 99%	Ridge	50,000	0.0184	stratified	-6e-05 (1.4e-05)	1.8e-07 (8.5e-09)	1.7e-07	0.00054	0.978	0.981
accuracy 99%	Ridge	100,000	0.0046	naive	-1.7e-05 (1e-05)	4e-08 (1.7e-09)	4e-08	0.00027	0.987	0.988
accuracy 99%	Ridge	100,000	0.0046	stratified	-1.8e-06 (1.1e-05)	4e-08 (1.8e-09)	4e-08	0.00028	0.987	0.988
accuracy 99%	Ridge	100,000	0.0092	naive	-2.2e-05 (1.1e-05)	6.6e-08 (2.8e-09)	6.5e-08	0.00034	0.984	0.986
accuracy 99%	Ridge	100,000	0.0092	stratified	-1.4e-05 (1e-05)	6.3e-08 (2.8e-09)	6.2e-08	0.00035	0.984	0.986
accuracy 99%	Ridge	100,000	0.0184	naive	-3e-05 (1e-05)	9e-08 (4.1e-09)	8.9e-08	0.0004	0.979	0.98
accuracy 99%	Ridge	100,000	0.0184	stratified	-3.2e-06 (1.1e-05)	8.8e-08 (4e-09)	8.9e-08	0.0004	0.978	0.98
accuracy 99%	Ridge	1,000,000	0.0046	naive	3.5e-06 (4.5e-06)	3.2e-09 (1.5e-10)	3.2e-09	7.3e-05	0.988	0.988
accuracy 99%	Ridge	1,000,000	0.0046	stratified	-9.2e-06 (4.3e-06)	3.3e-09 (1.5e-10)	3.3e-09	7.9e-05	0.988	0.988
accuracy 99%	Ridge	1,000,000	0.0092	naive	-7.1e-06 (4.3e-06)	3.9e-09 (1.8e-10)	3.9e-09	7.9e-05	0.985	0.986
accuracy 99%	Ridge	1,000,000	0.0092	stratified	-1.2e-05 (4.3e-06)	4.1e-09 (1.9e-10)	4e-09	8.4e-05	0.985	0.986

Table C.12: Bias, MSE, and Variability for accuracy at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 90%	Logistic	5,000	0.0046	naive	-0.18 (0.0094)	0.12 (0.0022)	0.083	0.45	0.0046	0.903
accuracy 90%	Logistic	5,000	0.0046	stratified	-0.17 (0.0095)	0.12 (0.0021)	0.088	0.45	0.0046	0.904
accuracy 90%	Logistic	5,000	0.0092	naive	-0.028 (0.0034)	0.028 (0.0034)	0.027	0.089	0.0092	0.909
accuracy 90%	Logistic	5,000	0.0092	stratified	-0.023 (0.0031)	0.025 (0.0033)	0.025	0.087	0.0092	0.907
accuracy 90%	Logistic	5,000	0.0184	naive	-0.0068 (0.0016)	0.0075 (0.0019)	0.0074	0.0046	0.0384	0.908
accuracy 90%	Logistic	5,000	0.0184	stratified	-0.0092 (0.002)	0.0083 (0.0018)	0.0082	0.0044	0.104	0.908
accuracy 90%	Logistic	10,000	0.0046	naive	-0.0044 (0.0012)	0.0017 (0.00073)	0.0017	0.0027	0.137	0.906
accuracy 90%	Logistic	10,000	0.0046	stratified	-0.0037 (0.0012)	0.0018 (0.0006)	0.0018	0.0029	0.307	0.907
accuracy 90%	Logistic	10,000	0.0092	naive	-0.00051 (0.00062)	0.00022 (0.00013)	0.00022	0.0027	0.544	0.905
accuracy 90%	Logistic	10,000	0.0092	stratified	-0.0013 (0.0005)	0.00025 (0.00021)	0.00025	0.0025	0.444	0.905
accuracy 90%	Logistic	10,000	0.0184	naive	-0.00042 (0.00014)	1.2e-05 (8.2e-06)	1.2e-05	0.0027	0.811	0.908
accuracy 90%	Logistic	10,000	0.0184	stratified	-0.00081 (0.00016)	2e-05 (1.7e-05)	2e-05	0.0025	0.773	0.907
accuracy 90%	Logistic	50,000	0.0046	naive	2.8e-05 (4.8e-05)	4.1e-07 (1.8e-08)	4.1e-07	0.00087	0.899	0.903
accuracy 90%	Logistic	50,000	0.0046	stratified	-3.1e-05 (4.8e-05)	4e-07 (1.9e-08)	4e-07	0.00086	0.898	0.903
accuracy 90%	Logistic	50,000	0.0092	naive	-4.4e-06 (4.7e-05)	4.1e-07 (1.8e-08)	4.1e-07	0.00089	0.9	0.904
accuracy 90%	Logistic	50,000	0.0092	stratified	-0.00011 (4.3e-05)	4.1e-07 (1.9e-08)	3.9e-07	0.00082	0.9	0.904
accuracy 90%	Logistic	50,000	0.0184	naive	-5.7e-05 (4.4e-05)	5.2e-07 (2.3e-08)	5.2e-07	0.00092	0.902	0.906
accuracy 90%	Logistic	50,000	0.0184	stratified	-0.00012 (4.6e-05)	5.5e-07 (2.4e-08)	5.4e-07	0.001	0.902	0.906
accuracy 90%	Logistic	100,000	0.0046	naive	-1.2e-05 (3.3e-05)	1.8e-07 (7.9e-09)	1.8e-07	0.00061	0.9	0.902
accuracy 90%	Logistic	100,000	0.0046	stratified	-5.3e-05 (3.2e-05)	1.8e-07 (8e-09)	1.8e-07	0.00059	0.9	0.902
accuracy 90%	Logistic	100,000	0.0092	naive	-9.4e-05 (3.2e-05)	1.7e-07 (8e-09)	1.7e-07	0.00053	0.901	0.904
accuracy 90%	Logistic	100,000	0.0092	stratified	3.2e-05 (3.3e-05)	1.7e-07 (7.5e-09)	1.7e-07	0.00056	0.901	0.903
accuracy 90%	Logistic	100,000	0.0184	naive	-4.4e-05 (3.3e-05)	2.2e-07 (9.6e-09)	2.2e-07	0.00065	0.903	0.906
accuracy 90%	Logistic	100,000	0.0184	stratified	-4.8e-05 (3.1e-05)	2.2e-07 (9.6e-09)	2.1e-07	0.00063	0.903	0.906
accuracy 90%	Logistic	1,000,000	0.0046	naive	-1.6e-05 (1.4e-05)	7.6e-09 (3.4e-10)	7.4e-09	0.00012	0.901	0.902
accuracy 90%	Logistic	1,000,000	0.0046	stratified	9.3e-06 (1.4e-05)	7.3e-09 (3.2e-10)	7.3e-09	0.00012	0.901	0.902
accuracy 90%	Logistic	1,000,000	0.0092	naive	1.2e-05 (1.4e-05)	8.5e-09 (3.7e-10)	8.3e-09	0.00013	0.902	0.903
accuracy 90%	Logistic	1,000,000	0.0092	stratified	-2.8e-06 (1.3e-05)	8.3e-09 (3.7e-10)	8.3e-09	0.00013	0.902	0.903

Table C.13: Bias, MSE, and Variability for accuracy at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 95%	Logistic	5,000	0.0046	naive	-0.19 (0.0099)	0.13 (0.0025)	0.095	0.51	0.0046	0.952
accuracy 95%	Logistic	5,000	0.0046	stratified	-0.18 (0.01)	0.13 (0.0024)	0.1	0.5	0.0046	0.952
accuracy 95%	Logistic	5,000	0.0092	naive	-0.029 (0.0036)	0.028 (0.0035)	0.028	0.093	0.0092	0.953
accuracy 95%	Logistic	5,000	0.0092	stratified	-0.022 (0.0033)	0.026 (0.0034)	0.025	0.092	0.0092	0.954
accuracy 95%	Logistic	5,000	0.0184	naive	-0.0074 (0.0018)	0.0077 (0.002)	0.0077	0.0038	0.0384	0.95
accuracy 95%	Logistic	5,000	0.0184	stratified	-0.0081 (0.0021)	0.0083 (0.0018)	0.0083	0.0036	0.11	0.952
accuracy 95%	Logistic	10,000	0.0046	naive	-0.0048 (0.0013)	0.002 (0.00082)	0.0019	0.0022	0.14	0.953
accuracy 95%	Logistic	10,000	0.0046	stratified	-0.0032 (0.0012)	0.0017 (0.00056)	0.0017	0.0022	0.368	0.953
accuracy 95%	Logistic	10,000	0.0092	naive	-0.0007 (0.00065)	0.00024 (0.00015)	0.00024	0.002	0.57	0.952
accuracy 95%	Logistic	10,000	0.0092	stratified	-0.0013 (0.00053)	0.00029 (0.00024)	0.00029	0.002	0.459	0.952
accuracy 95%	Logistic	10,000	0.0184	naive	-0.00055 (0.00012)	1.2e-05 (8.5e-06)	1.1e-05	0.0022	0.854	0.951
accuracy 95%	Logistic	10,000	0.0184	stratified	-0.00063 (0.00016)	2.2e-05 (1.9e-05)	2.2e-05	0.0022	0.809	0.951
accuracy 95%	Logistic	50,000	0.0046	naive	-2.6e-05 (3.4e-05)	2.6e-07 (1.2e-08)	2.6e-07	0.00068	0.948	0.951
accuracy 95%	Logistic	50,000	0.0046	stratified	-7.9e-05 (3.3e-05)	2.7e-07 (1.2e-08)	2.7e-07	0.00068	0.948	0.951
accuracy 95%	Logistic	50,000	0.0092	naive	-7.3e-05 (3.3e-05)	3.2e-07 (1.5e-08)	3.2e-07	0.00077	0.947	0.951
accuracy 95%	Logistic	50,000	0.0092	stratified	-9.4e-05 (3.2e-05)	3.1e-07 (1.4e-08)	3e-07	0.00074	0.948	0.951
accuracy 95%	Logistic	50,000	0.0184	naive	-7.8e-05 (3.1e-05)	4.2e-07 (1.8e-08)	4.1e-07	0.0009	0.947	0.951
accuracy 95%	Logistic	50,000	0.0184	stratified	-9.4e-05 (3.3e-05)	4.4e-07 (2e-08)	4.3e-07	0.00089	0.947	0.951
accuracy 95%	Logistic	100,000	0.0046	naive	-5.1e-05 (2.5e-05)	1.2e-07 (5.1e-09)	1.2e-07	0.00048	0.949	0.951
accuracy 95%	Logistic	100,000	0.0046	stratified	-5.6e-05 (2.4e-05)	1.2e-07 (5.3e-09)	1.1e-07	0.00044	0.949	0.951
accuracy 95%	Logistic	100,000	0.0092	naive	-6.2e-05 (2.4e-05)	1.4e-07 (6.5e-09)	1.4e-07	0.00049	0.948	0.951
accuracy 95%	Logistic	100,000	0.0092	stratified	-3.4e-05 (2.4e-05)	1.5e-07 (6.6e-09)	1.5e-07	0.00052	0.949	0.951
accuracy 95%	Logistic	100,000	0.0184	naive	-2.6e-05 (2.3e-05)	2e-07 (8.2e-09)	2e-07	0.00059	0.948	0.951
accuracy 95%	Logistic	100,000	0.0184	stratified	-6.1e-05 (2.3e-05)	1.8e-07 (7.7e-09)	1.8e-07	0.00059	0.948	0.95
accuracy 95%	Logistic	1,000,000	0.0046	naive	-9.1e-06 (1e-05)	6.2e-09 (2.8e-10)	6.1e-09	0.0001	0.95	0.95
accuracy 95%	Logistic	1,000,000	0.0046	stratified	9.6e-06 (1e-05)	6.3e-09 (2.6e-10)	6.2e-09	0.00011	0.95	0.95
accuracy 95%	Logistic	1,000,000	0.0092	naive	-2.7e-07 (1e-05)	7.5e-09 (3.6e-10)	7.5e-09	0.00011	0.95	0.95
accuracy 95%	Logistic	1,000,000	0.0092	stratified	-8.2e-06 (9.9e-06)	7.9e-09 (3.6e-10)	7.9e-09	0.00012	0.95	0.95

Table C.14: Bias, MSE, and Variability for accuracy at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
accuracy 99%	Logistic	5,000	0.0046	naive	-0.025 (0.013)	0.061 (0.0024)	0.061	0.38	0.0046	0.984
accuracy 99%	Logistic	5,000	0.0046	stratified	-0.012 (0.013)	0.065 (0.0023)	0.065	0.4	0.0046	0.984
accuracy 99%	Logistic	5,000	0.0092	naive	-0.026 (0.0037)	0.027 (0.0033)	0.026	0.096	0.0092	0.982
accuracy 99%	Logistic	5,000	0.0092	stratified	-0.023 (0.0035)	0.025 (0.0035)	0.024	0.095	0.0092	0.984
accuracy 99%	Logistic	5,000	0.0184	naive	-0.0046 (0.0019)	0.0038 (0.0011)	0.0038	0.0024	0.208	0.979
accuracy 99%	Logistic	5,000	0.0184	stratified	-0.0037 (0.0022)	0.0048 (0.0011)	0.0048	0.0022	0.326	0.979
accuracy 99%	Logistic	10,000	0.0046	naive	-0.0044 (0.0013)	0.0015 (0.00051)	0.0015	0.0012	0.433	0.988
accuracy 99%	Logistic	10,000	0.0046	stratified	-0.0029 (0.0013)	0.0016 (0.00057)	0.0016	0.0014	0.376	0.988
accuracy 99%	Logistic	10,000	0.0092	naive	-0.00066 (0.00065)	0.00023 (0.00015)	0.00023	0.0011	0.593	0.985
accuracy 99%	Logistic	10,000	0.0092	stratified	-0.0013 (0.00055)	0.00031 (0.00026)	0.00031	0.0013	0.469	0.985
accuracy 99%	Logistic	10,000	0.0184	naive	-0.00032 (3.9e-05)	1.5e-06 (2.9e-07)	1.4e-06	0.0015	0.961	0.981
accuracy 99%	Logistic	10,000	0.0184	stratified	-0.00052 (0.00015)	2.3e-05 (2.1e-05)	2.2e-05	0.0015	0.831	0.981
accuracy 99%	Logistic	50,000	0.0046	naive	-5.4e-05 (1.6e-05)	1.1e-07 (4.7e-09)	1e-07	0.00044	0.986	0.988
accuracy 99%	Logistic	50,000	0.0046	stratified	-6.3e-05 (1.6e-05)	1e-07 (4.8e-09)	1e-07	0.00042	0.986	0.989
accuracy 99%	Logistic	50,000	0.0092	naive	-6e-05 (1.5e-05)	1.4e-07 (6.3e-09)	1.4e-07	0.0005	0.984	0.986
accuracy 99%	Logistic	50,000	0.0092	stratified	-9.1e-05 (1.5e-05)	1.5e-07 (6.3e-09)	1.4e-07	0.00052	0.984	0.986
accuracy 99%	Logistic	50,000	0.0184	naive	-6e-05 (1.4e-05)	1.9e-07 (8.5e-09)	1.9e-07	0.0006	0.977	0.981
accuracy 99%	Logistic	50,000	0.0184	stratified	-5.8e-05 (1.5e-05)	2.1e-07 (9.3e-09)	2e-07	0.0006	0.978	0.981
accuracy 99%	Logistic	100,000	0.0046	naive	-2.8e-05 (1.1e-05)	4.7e-08 (2.1e-09)	4.6e-08	0.00029	0.987	0.988
accuracy 99%	Logistic	100,000	0.0046	stratified	-4.3e-05 (1.1e-05)	4.7e-08 (2.1e-09)	4.6e-08	0.00028	0.987	0.988
accuracy 99%	Logistic	100,000	0.0092	naive	-4.6e-05 (1.1e-05)	7.1e-08 (3.3e-09)	6.9e-08	0.00035	0.984	0.986
accuracy 99%	Logistic	100,000	0.0092	stratified	-4.4e-05 (1.1e-05)	7e-08 (3.1e-09)	6.8e-08	0.00034	0.984	0.986
accuracy 99%	Logistic	100,000	0.0184	naive	-1.7e-05 (1e-05)	9e-08 (4.1e-09)	9e-08	0.00041	0.978	0.981
accuracy 99%	Logistic	100,000	0.0184	stratified	-3.2e-05 (1e-05)	8.5e-08 (3.8e-09)	8.4e-08	0.00036	0.978	0.98
accuracy 99%	Logistic	1,000,000	0.0046	naive	-4.6e-07 (4.5e-06)	3.4e-09 (1.5e-10)	3.4e-09	7.6e-05	0.988	0.988
accuracy 99%	Logistic	1,000,000	0.0046	stratified	-2.9e-06 (4.7e-06)	3.4e-09 (1.6e-10)	3.4e-09	7.5e-05	0.988	0.988
accuracy 99%	Logistic	1,000,000	0.0092	naive	-1.7e-06 (4.3e-06)	4.7e-09 (2e-10)	4.7e-09	9.7e-05	0.985	0.986
accuracy 99%	Logistic	1,000,000	0.0092	stratified	3.7e-06 (4.3e-06)	4.5e-09 (2e-10)	4.5e-09	8.8e-05	0.985	0.985

Table C.15: Bias, MSE, and Variability for accuracy at the 99th percentile, logistic regression.

C.4 Sensitivity

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 90%	Random Forest	5,000	0.0046	naive	-0.0063 (0.0058)	0.018 (0.001)	0.018	0.19	0.185	0.922
sensitivity 90%	Random Forest	5,000	0.0046	stratified	-0.0067 (0.0049)	0.013 (0.0008)	0.013	0.15	0.2	0.833
sensitivity 90%	Random Forest	5,000	0.0092	naive	-0.002 (0.004)	0.0076 (0.00046)	0.0076	0.12	0.25	0.799
sensitivity 90%	Random Forest	5,000	0.0092	stratified	-0.0054 (0.0033)	0.0057 (0.00035)	0.0057	0.1	0.365	0.8
sensitivity 90%	Random Forest	5,000	0.0184	naive	-0.0018 (0.0025)	0.0028 (0.00017)	0.0028	0.074	0.424	0.749
sensitivity 90%	Random Forest	5,000	0.0184	stratified	-0.0012 (0.0024)	0.0027 (0.00016)	0.0027	0.069	0.42	0.739
sensitivity 90%	Random Forest	10,000	0.0046	naive	0.0042 (0.0039)	0.0068 (0.00038)	0.0068	0.12	0.331	0.796
sensitivity 90%	Random Forest	10,000	0.0046	stratified	0.00068 (0.0034)	0.0053 (0.00033)	0.0053	0.1	0.34	0.785
sensitivity 90%	Random Forest	10,000	0.0092	naive	-0.0023 (0.0027)	0.0035 (0.00021)	0.0035	0.079	0.437	0.769
sensitivity 90%	Random Forest	10,000	0.0092	stratified	0.0012 (0.0024)	0.0027 (0.00016)	0.0027	0.067	0.458	0.729
sensitivity 90%	Random Forest	10,000	0.0184	naive	-0.004 (0.0018)	0.0015 (8.7e-05)	0.0015	0.056	0.501	0.712
sensitivity 90%	Random Forest	10,000	0.0184	stratified	0.002 (0.0016)	0.0013 (7.8e-05)	0.0013	0.049	0.511	0.712
sensitivity 90%	Random Forest	50,000	0.0046	naive	-0.0015 (0.0015)	0.001 (6.8e-05)	0.001	0.045	0.493	0.698
sensitivity 90%	Random Forest	50,000	0.0046	stratified	-0.0021 (0.0015)	0.00094 (5.8e-05)	0.00094	0.039	0.504	0.7
sensitivity 90%	Random Forest	50,000	0.0092	naive	-0.0032 (0.001)	0.00048 (3e-05)	0.00047	0.029	0.541	0.682
sensitivity 90%	Random Forest	50,000	0.0092	stratified	-0.0027 (0.0011)	0.00051 (3.2e-05)	0.0005	0.03	0.546	0.685
sensitivity 90%	Random Forest	50,000	0.0184	naive	-0.0037 (0.00075)	0.00025 (1.7e-05)	0.00024	0.021	0.555	0.649
sensitivity 90%	Random Forest	50,000	0.0184	stratified	-0.0018 (0.00077)	0.00024 (1.5e-05)	0.00024	0.02	0.561	0.652
sensitivity 90%	Random Forest	100,000	0.0046	naive	-0.0024 (0.0011)	0.00051 (3.3e-05)	0.0005	0.03	0.552	0.695
sensitivity 90%	Random Forest	100,000	0.0046	stratified	-0.0016 (0.0011)	0.00052 (3.2e-05)	0.00052	0.028	0.563	0.696
sensitivity 90%	Random Forest	100,000	0.0092	naive	-0.0014 (0.00073)	0.00022 (1.3e-05)	0.00021	0.02	0.581	0.66
sensitivity 90%	Random Forest	100,000	0.0092	stratified	-0.00077 (0.00078)	0.00025 (1.6e-05)	0.00025	0.021	0.579	0.683
sensitivity 90%	Random Forest	100,000	0.0184	naive	-0.0014 (0.00056)	0.00012 (7.3e-06)	0.00012	0.015	0.584	0.645
sensitivity 90%	Random Forest	100,000	0.0184	stratified	-0.0021 (0.00056)	0.00012 (7.2e-06)	0.00011	0.015	0.585	0.648
sensitivity 90%	Random Forest	1,000,000	0.0046	naive	-0.0013 (0.00046)	4.2e-05 (2.6e-06)	4.1e-05	0.0088	0.613	0.651
sensitivity 90%	Random Forest	1,000,000	0.0046	stratified	-0.0013 (0.00044)	4.1e-05 (2.6e-06)	3.9e-05	0.0081	0.608	0.652
sensitivity 90%	Random Forest	1,000,000	0.0092	naive	-0.0014 (0.00031)	1.7e-05 (1e-06)	1.5e-05	0.0053	0.614	0.637
sensitivity 90%	Random Forest	1,000,000	0.0092	stratified	-0.0018 (0.00034)	1.8e-05 (1e-06)	1.5e-05	0.0052	0.616	0.636

Table C.16: Bias, MSE, and Variability for sensitivity at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 95%	Random Forest	5,000	0.0046	naive	-0.01 (0.0057)	0.017 (0.001)	0.017	0.18	0.0833	0.806
sensitivity 95%	Random Forest	5,000	0.0046	stratified	-0.0077 (0.005)	0.013 (0.00081)	0.013	0.17	0.05	0.75
sensitivity 95%	Random Forest	5,000	0.0092	naive	-0.0032 (0.0039)	0.0075 (0.00048)	0.0075	0.11	0.171	0.672
sensitivity 95%	Random Forest	5,000	0.0092	stratified	-0.002 (0.0035)	0.006 (0.0004)	0.006	0.1	0.15	0.69
sensitivity 95%	Random Forest	5,000	0.0184	naive	-7.8e-05 (0.0025)	0.0028 (0.00017)	0.0028	0.075	0.293	0.613
sensitivity 95%	Random Forest	5,000	0.0184	stratified	-0.0024 (0.0024)	0.0026 (0.00016)	0.0026	0.069	0.282	0.587
sensitivity 95%	Random Forest	10,000	0.0046	naive	0.0054 (0.0038)	0.0065 (0.00038)	0.0065	0.11	0.206	0.651
sensitivity 95%	Random Forest	10,000	0.0046	stratified	-0.0031 (0.0034)	0.0053 (0.00032)	0.0053	0.1	0.22	0.63
sensitivity 95%	Random Forest	10,000	0.0092	naive	-0.0053 (0.0026)	0.0035 (0.00023)	0.0035	0.073	0.283	0.676
sensitivity 95%	Random Forest	10,000	0.0092	stratified	0.0011 (0.0024)	0.0026 (0.00018)	0.0026	0.067	0.292	0.631
sensitivity 95%	Random Forest	10,000	0.0184	naive	-0.0041 (0.0017)	0.0012 (7.9e-05)	0.0012	0.044	0.354	0.571
sensitivity 95%	Random Forest	10,000	0.0184	stratified	-0.00043 (0.0017)	0.0012 (7.8e-05)	0.0012	0.045	0.358	0.567
sensitivity 95%	Random Forest	50,000	0.0046	naive	-0.0019 (0.0015)	0.001 (6.4e-05)	0.001	0.039	0.379	0.571
sensitivity 95%	Random Forest	50,000	0.0046	stratified	-0.00074 (0.0015)	0.00098 (6.3e-05)	0.00098	0.039	0.387	0.57
sensitivity 95%	Random Forest	50,000	0.0092	naive	-0.002 (0.0011)	0.00051 (3.1e-05)	0.00051	0.031	0.415	0.538
sensitivity 95%	Random Forest	50,000	0.0092	stratified	-0.0027 (0.0012)	0.00055 (3.6e-05)	0.00054	0.03	0.402	0.548
sensitivity 95%	Random Forest	50,000	0.0184	naive	-0.0037 (0.00077)	0.00025 (1.6e-05)	0.00024	0.02	0.419	0.511
sensitivity 95%	Random Forest	50,000	0.0184	stratified	-0.003 (0.00081)	0.00027 (1.6e-05)	0.00026	0.02	0.416	0.509
sensitivity 95%	Random Forest	100,000	0.0046	naive	-0.0023 (0.0011)	0.00054 (3.7e-05)	0.00053	0.027	0.415	0.555
sensitivity 95%	Random Forest	100,000	0.0046	stratified	-0.0026 (0.0011)	0.00057 (3.4e-05)	0.00056	0.033	0.426	0.567
sensitivity 95%	Random Forest	100,000	0.0092	naive	-0.0013 (0.00072)	0.00021 (1.3e-05)	0.00021	0.02	0.445	0.525
sensitivity 95%	Random Forest	100,000	0.0092	stratified	-0.00098 (0.00077)	0.00024 (1.5e-05)	0.00024	0.021	0.428	0.532
sensitivity 95%	Random Forest	100,000	0.0184	naive	-0.0016 (0.00057)	0.00012 (8e-06)	0.00012	0.014	0.435	0.511
sensitivity 95%	Random Forest	100,000	0.0184	stratified	-0.0016 (0.0006)	0.00013 (7.4e-06)	0.00013	0.016	0.442	0.505
sensitivity 95%	Random Forest	1,000,000	0.0046	naive	-0.0014 (0.00047)	4.6e-05 (2.8e-06)	4.5e-05	0.009	0.474	0.513
sensitivity 95%	Random Forest	1,000,000	0.0046	stratified	-0.0019 (0.00044)	4.1e-05 (2.3e-06)	3.7e-05	0.0081	0.477	0.513
sensitivity 95%	Random Forest	1,000,000	0.0092	naive	-0.001 (0.00034)	1.9e-05 (1.2e-06)	1.8e-05	0.0058	0.473	0.498
sensitivity 95%	Random Forest	1,000,000	0.0092	stratified	-0.0012 (0.00033)	1.8e-05 (1.1e-06)	1.6e-05	0.0055	0.476	0.499

Table C.17: Bias, MSE, and Variability for sensitivity at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 99%	Random Forest	5,000	0.0046	naive	-0.0013 (0.0045)	0.011 (0.00065)	0.011	0.15	0	0.537
sensitivity 99%	Random Forest	5,000	0.0046	stratified	-0.0017 (0.0041)	0.0093 (0.00052)	0.0093	0.12	0	0.45
sensitivity 99%	Random Forest	5,000	0.0092	naive	-0.0021 (0.0029)	0.0044 (0.0003)	0.0044	0.087	0.02	0.416
sensitivity 99%	Random Forest	5,000	0.0092	stratified	0.0002 (0.0029)	0.0043 (0.00024)	0.0043	0.095	0.025	0.35
sensitivity 99%	Random Forest	5,000	0.0184	naive	-0.0022 (0.002)	0.0017 (0.00011)	0.0017	0.056	0.0361	0.301
sensitivity 99%	Random Forest	5,000	0.0184	stratified	-0.0012 (0.0018)	0.0014 (9.5e-05)	0.0014	0.047	0.0444	0.293
sensitivity 99%	Random Forest	10,000	0.0046	naive	0.00029 (0.0032)	0.005 (0.00029)	0.0051	0.093	0.025	0.378
sensitivity 99%	Random Forest	10,000	0.0046	stratified	-0.0038 (0.0029)	0.0043 (0.00027)	0.0043	0.09	0	0.405
sensitivity 99%	Random Forest	10,000	0.0092	naive	-0.0041 (0.002)	0.0021 (0.00013)	0.0021	0.06	0.0624	0.337
sensitivity 99%	Random Forest	10,000	0.0092	stratified	0.0012 (0.002)	0.002 (0.00011)	0.002	0.063	0.0667	0.327
sensitivity 99%	Random Forest	10,000	0.0184	naive	-0.0027 (0.0014)	0.00079 (5.5e-05)	0.00079	0.038	0.0972	0.298
sensitivity 99%	Random Forest	10,000	0.0184	stratified	-0.0017 (0.0013)	0.00076 (5.1e-05)	0.00076	0.037	0.0813	0.268
sensitivity 99%	Random Forest	50,000	0.0046	naive	-0.00086 (0.0014)	0.00081 (5.2e-05)	0.00081	0.038	0.14	0.308
sensitivity 99%	Random Forest	50,000	0.0046	stratified	-0.0011 (0.0013)	0.00075 (4.6e-05)	0.00075	0.035	0.148	0.322
sensitivity 99%	Random Forest	50,000	0.0092	naive	-0.00068 (0.00088)	0.0003 (1.9e-05)	0.0003	0.023	0.162	0.276
sensitivity 99%	Random Forest	50,000	0.0092	stratified	-0.0017 (0.0009)	0.00028 (1.8e-05)	0.00027	0.02	0.176	0.278
sensitivity 99%	Random Forest	50,000	0.0184	naive	-0.0015 (0.00061)	0.00011 (6.8e-06)	0.00011	0.014	0.171	0.233
sensitivity 99%	Random Forest	50,000	0.0184	stratified	-0.0011 (0.00062)	0.00011 (7.6e-06)	0.00011	0.014	0.165	0.24
sensitivity 99%	Random Forest	100,000	0.0046	naive	-0.00048 (0.00095)	0.00036 (2.1e-05)	0.00036	0.026	0.181	0.288
sensitivity 99%	Random Forest	100,000	0.0046	stratified	-0.0017 (0.00094)	0.00035 (2.4e-05)	0.00035	0.024	0.174	0.285
sensitivity 99%	Random Forest	100,000	0.0092	naive	-0.0011 (0.00065)	0.00015 (8.8e-06)	0.00014	0.016	0.188	0.263
sensitivity 99%	Random Forest	100,000	0.0092	stratified	0.00077 (0.00064)	0.00014 (8.6e-06)	0.00014	0.015	0.194	0.263
sensitivity 99%	Random Forest	100,000	0.0184	naive	-0.0006 (0.00043)	5.2e-05 (3.4e-06)	5.2e-05	0.01	0.183	0.229
sensitivity 99%	Random Forest	100,000	0.0184	stratified	-0.0013 (0.00046)	5.7e-05 (3.7e-06)	5.5e-05	0.01	0.178	0.221
sensitivity 99%	Random Forest	1,000,000	0.0046	naive	-0.00092 (0.00042)	3e-05 (2e-06)	2.9e-05	0.0068	0.219	0.253
sensitivity 99%	Random Forest	1,000,000	0.0046	stratified	-0.00053 (0.00039)	2.4e-05 (1.5e-06)	2.4e-05	0.0065	0.223	0.249
sensitivity 99%	Random Forest	1,000,000	0.0092	naive	-0.00058 (0.00029)	1e-05 (6.1e-07)	1e-05	0.0044	0.214	0.231
sensitivity 99%	Random Forest	1,000,000	0.0092	stratified	-0.00048 (0.00029)	9.8e-06 (6.4e-07)	9.6e-06	0.0041	0.214	0.233

Table C.18: Bias, MSE, and Variability for sensitivity at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 90%	Ridge	5,000	0.0046	naive	-2.3e-05 (0.0038)	0.015 (0.00064)	0.015	0.18	0.2	0.88
sensitivity 90%	Ridge	5,000	0.0046	stratified	-0.0023 (0.0035)	0.013 (0.0006)	0.013	0.15	0.05	0.967
sensitivity 90%	Ridge	5,000	0.0092	naive	0.0022 (0.0027)	0.0069 (0.0003)	0.0069	0.11	0.31	0.794
sensitivity 90%	Ridge	5,000	0.0092	stratified	-0.0022 (0.0024)	0.0056 (0.00025)	0.0056	0.1	0.335	0.83
sensitivity 90%	Ridge	5,000	0.0184	naive	-0.00023 (0.0018)	0.0029 (0.00013)	0.0029	0.073	0.401	0.74
sensitivity 90%	Ridge	5,000	0.0184	stratified	-0.0026 (0.0017)	0.0025 (0.00011)	0.0025	0.072	0.403	0.74
sensitivity 90%	Ridge	10,000	0.0046	naive	-0.00021 (0.0028)	0.0074 (0.0003)	0.0074	0.13	0.325	0.836
sensitivity 90%	Ridge	10,000	0.0046	stratified	-0.0035 (0.0023)	0.0052 (0.00024)	0.0052	0.095	0.33	0.805
sensitivity 90%	Ridge	10,000	0.0092	naive	-0.0039 (0.0018)	0.0029 (0.00013)	0.0029	0.074	0.424	0.78
sensitivity 90%	Ridge	10,000	0.0092	stratified	-0.0027 (0.0017)	0.0028 (0.00012)	0.0028	0.068	0.422	0.751
sensitivity 90%	Ridge	10,000	0.0184	naive	-0.003 (0.0012)	0.0014 (5.9e-05)	0.0014	0.051	0.456	0.685
sensitivity 90%	Ridge	10,000	0.0184	stratified	-0.0021 (0.0012)	0.0014 (6.1e-05)	0.0014	0.051	0.472	0.719
sensitivity 90%	Ridge	50,000	0.0046	naive	-0.0022 (0.0011)	0.0012 (5.4e-05)	0.0012	0.048	0.499	0.721
sensitivity 90%	Ridge	50,000	0.0046	stratified	-0.0034 (0.001)	0.001 (4.7e-05)	0.001	0.043	0.487	0.717
sensitivity 90%	Ridge	50,000	0.0092	naive	-0.0014 (0.00073)	0.0005 (2.2e-05)	0.00049	0.03	0.543	0.686
sensitivity 90%	Ridge	50,000	0.0092	stratified	-0.0019 (0.00075)	0.0005 (2.3e-05)	0.0005	0.033	0.537	0.687
sensitivity 90%	Ridge	50,000	0.0184	naive	-0.0013 (0.00053)	0.00025 (1.2e-05)	0.00025	0.02	0.565	0.67
sensitivity 90%	Ridge	50,000	0.0184	stratified	-0.00076 (0.00053)	0.00025 (1.1e-05)	0.00025	0.023	0.56	0.664
sensitivity 90%	Ridge	100,000	0.0046	naive	-0.0018 (0.00077)	0.00053 (2.4e-05)	0.00052	0.03	0.553	0.7
sensitivity 90%	Ridge	100,000	0.0046	stratified	-0.00065 (0.00079)	0.00057 (2.6e-05)	0.00057	0.03	0.55	0.691
sensitivity 90%	Ridge	100,000	0.0092	naive	-0.0014 (0.00054)	0.00025 (1.1e-05)	0.00025	0.021	0.575	0.677
sensitivity 90%	Ridge	100,000	0.0092	stratified	-0.00053 (0.00054)	0.00025 (1.1e-05)	0.00025	0.021	0.576	0.674
sensitivity 90%	Ridge	100,000	0.0184	naive	-0.001 (0.00038)	0.00011 (5e-06)	0.00011	0.015	0.59	0.661
sensitivity 90%	Ridge	100,000	0.0184	stratified	-0.00049 (0.00038)	0.00012 (5.6e-06)	0.00012	0.014	0.588	0.667
sensitivity 90%	Ridge	1,000,000	0.0046	naive	3.8e-05 (0.00032)	4.4e-05 (1.9e-06)	4.4e-05	0.0086	0.621	0.665
sensitivity 90%	Ridge	1,000,000	0.0046	stratified	-0.00049 (0.00032)	4.4e-05 (1.8e-06)	4.4e-05	0.0098	0.623	0.661
sensitivity 90%	Ridge	1,000,000	0.0092	naive	-0.00019 (0.00022)	1.6e-05 (7.3e-07)	1.6e-05	0.0051	0.627	0.652
sensitivity 90%	Ridge	1,000,000	0.0092	stratified	-0.00036 (0.00022)	1.6e-05 (6.9e-07)	1.6e-05	0.0056	0.625	0.651

Table C.19: Bias, MSE, and Variability for sensitivity at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 95%	Ridge	5,000	0.0046	naive	0.00086 (0.0039)	0.016 (0.00064)	0.016	0.18	0.09	0.815
sensitivity 95%	Ridge	5,000	0.0046	stratified	-0.0054 (0.0035)	0.013 (0.00058)	0.013	0.15	0	0.783
sensitivity 95%	Ridge	5,000	0.0092	naive	0.0011 (0.0027)	0.007 (0.0003)	0.007	0.12	0.156	0.692
sensitivity 95%	Ridge	5,000	0.0092	stratified	-0.0013 (0.0024)	0.0057 (0.00026)	0.0057	0.1	0.2	0.69
sensitivity 95%	Ridge	5,000	0.0184	naive	4.7e-05 (0.0018)	0.0029 (0.00013)	0.0029	0.074	0.243	0.633
sensitivity 95%	Ridge	5,000	0.0184	stratified	-0.0021 (0.0017)	0.0026 (0.00011)	0.0026	0.069	0.272	0.577
sensitivity 95%	Ridge	10,000	0.0046	naive	-0.00094 (0.0028)	0.0075 (0.00033)	0.0075	0.11	0.171	0.717
sensitivity 95%	Ridge	10,000	0.0046	stratified	-0.0035 (0.0024)	0.0054 (0.00025)	0.0054	0.1	0.21	0.7
sensitivity 95%	Ridge	10,000	0.0092	naive	-0.0037 (0.0018)	0.003 (0.00014)	0.003	0.076	0.259	0.615
sensitivity 95%	Ridge	10,000	0.0092	stratified	-0.0038 (0.0017)	0.0028 (0.00012)	0.0028	0.068	0.293	0.619
sensitivity 95%	Ridge	10,000	0.0184	naive	-0.0027 (0.0013)	0.0015 (6.4e-05)	0.0015	0.053	0.329	0.585
sensitivity 95%	Ridge	10,000	0.0184	stratified	-0.0011 (0.0012)	0.0013 (5.5e-05)	0.0013	0.048	0.348	0.57
sensitivity 95%	Ridge	50,000	0.0046	naive	-0.002 (0.0011)	0.0012 (5.3e-05)	0.0012	0.045	0.351	0.582
sensitivity 95%	Ridge	50,000	0.0046	stratified	-0.0037 (0.0011)	0.0011 (4.9e-05)	0.0011	0.043	0.37	0.591
sensitivity 95%	Ridge	50,000	0.0092	naive	-0.0017 (0.00075)	0.00052 (2.4e-05)	0.00052	0.029	0.392	0.543
sensitivity 95%	Ridge	50,000	0.0092	stratified	-0.0011 (0.00076)	0.0005 (2.2e-05)	0.0005	0.033	0.411	0.546
sensitivity 95%	Ridge	50,000	0.0184	naive	-0.0012 (0.00055)	0.00027 (1.3e-05)	0.00026	0.021	0.416	0.533
sensitivity 95%	Ridge	50,000	0.0184	stratified	-0.00095 (0.00053)	0.00024 (1.1e-05)	0.00024	0.021	0.429	0.536
sensitivity 95%	Ridge	100,000	0.0046	naive	-0.0015 (0.0008)	0.00057 (2.5e-05)	0.00057	0.033	0.412	0.565
sensitivity 95%	Ridge	100,000	0.0046	stratified	-0.00099 (0.00081)	0.00058 (2.6e-05)	0.00058	0.033	0.413	0.557
sensitivity 95%	Ridge	100,000	0.0092	naive	-0.0016 (0.00055)	0.00027 (1.2e-05)	0.00026	0.022	0.437	0.544
sensitivity 95%	Ridge	100,000	0.0092	stratified	-0.0011 (0.00054)	0.00026 (1.1e-05)	0.00026	0.021	0.441	0.54
sensitivity 95%	Ridge	100,000	0.0184	naive	-0.00085 (0.00038)	0.00011 (4.8e-06)	0.00011	0.014	0.445	0.51
sensitivity 95%	Ridge	100,000	0.0184	stratified	-0.00067 (0.0004)	0.00013 (6.2e-06)	0.00013	0.015	0.441	0.531
sensitivity 95%	Ridge	1,000,000	0.0046	naive	4.5e-05 (0.00033)	4.5e-05 (2e-06)	4.6e-05	0.0089	0.49	0.53
sensitivity 95%	Ridge	1,000,000	0.0046	stratified	-0.00022 (0.00033)	4.8e-05 (2.1e-06)	4.8e-05	0.0094	0.489	0.532
sensitivity 95%	Ridge	1,000,000	0.0092	naive	-0.00031 (0.00023)	1.7e-05 (7.5e-07)	1.7e-05	0.0056	0.491	0.517
sensitivity 95%	Ridge	1,000,000	0.0092	stratified	-0.00013 (0.00022)	1.7e-05 (7.8e-07)	1.7e-05	0.0053	0.488	0.517

Table C.20: Bias, MSE, and Variability for sensitivity at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 99%	Ridge	5,000	0.0046	naive	0.0028 (0.0032)	0.011 (0.00045)	0.011	0.15	0	0.574
sensitivity 99%	Ridge	5,000	0.0046	stratified	-0.006 (0.0029)	0.0089 (0.00036)	0.0088	0.13	0	0.517
sensitivity 99%	Ridge	5,000	0.0092	naive	-0.0011 (0.0022)	0.0048 (0.0002)	0.0048	0.095	0.02	0.426
sensitivity 99%	Ridge	5,000	0.0092	stratified	-0.0019 (0.002)	0.0039 (0.00017)	0.0039	0.08	0.025	0.425
sensitivity 99%	Ridge	5,000	0.0184	naive	0.0026 (0.0014)	0.0016 (7.5e-05)	0.0016	0.057	0.0665	0.326
sensitivity 99%	Ridge	5,000	0.0184	stratified	-0.0019 (0.0013)	0.0015 (6.5e-05)	0.0015	0.053	0.0744	0.316
sensitivity 99%	Ridge	10,000	0.0046	naive	-5.3e-05 (0.0022)	0.0047 (0.00021)	0.0047	0.092	0.0393	0.444
sensitivity 99%	Ridge	10,000	0.0046	stratified	-0.0022 (0.0021)	0.0043 (0.00019)	0.0043	0.09	0.04	0.415
sensitivity 99%	Ridge	10,000	0.0092	naive	-0.0015 (0.0016)	0.0022 (9.3e-05)	0.0022	0.069	0.087	0.387
sensitivity 99%	Ridge	10,000	0.0092	stratified	-0.00035 (0.0015)	0.002 (8.7e-05)	0.002	0.057	0.0867	0.348
sensitivity 99%	Ridge	10,000	0.0184	naive	0.00047 (0.001)	0.0008 (3.7e-05)	0.0008	0.038	0.102	0.29
sensitivity 99%	Ridge	10,000	0.0184	stratified	-0.0017 (0.00098)	0.00076 (3.2e-05)	0.00075	0.038	0.12	0.278
sensitivity 99%	Ridge	50,000	0.0046	naive	-0.002 (0.00094)	0.00077 (3.6e-05)	0.00077	0.037	0.153	0.344
sensitivity 99%	Ridge	50,000	0.0046	stratified	-0.004 (0.00092)	0.00079 (3.3e-05)	0.00077	0.039	0.148	0.317
sensitivity 99%	Ridge	50,000	0.0092	naive	-0.00071 (0.00063)	0.00033 (1.5e-05)	0.00033	0.024	0.171	0.282
sensitivity 99%	Ridge	50,000	0.0092	stratified	-0.00064 (0.00067)	0.00036 (1.6e-05)	0.00036	0.026	0.172	0.283
sensitivity 99%	Ridge	50,000	0.0184	naive	-0.0011 (0.00043)	0.00014 (6e-06)	0.00014	0.016	0.175	0.247
sensitivity 99%	Ridge	50,000	0.0184	stratified	-0.0011 (0.00043)	0.00013 (6.2e-06)	0.00013	0.015	0.168	0.247
sensitivity 99%	Ridge	100,000	0.0046	naive	-0.00095 (0.00067)	0.00038 (1.6e-05)	0.00038	0.027	0.183	0.296
sensitivity 99%	Ridge	100,000	0.0046	stratified	-0.00054 (0.00067)	0.00039 (1.6e-05)	0.00039	0.026	0.191	0.3
sensitivity 99%	Ridge	100,000	0.0092	naive	-0.0016 (0.00047)	0.00017 (7.5e-06)	0.00017	0.018	0.196	0.277
sensitivity 99%	Ridge	100,000	0.0092	stratified	-0.0011 (0.00047)	0.00017 (7.6e-06)	0.00017	0.018	0.194	0.28
sensitivity 99%	Ridge	100,000	0.0184	naive	-0.0003 (0.00032)	6.4e-05 (2.9e-06)	6.4e-05	0.011	0.188	0.24
sensitivity 99%	Ridge	100,000	0.0184	stratified	-0.00065 (0.00032)	6.5e-05 (3e-06)	6.5e-05	0.011	0.183	0.236
sensitivity 99%	Ridge	1,000,000	0.0046	naive	-0.00024 (0.0003)	3.4e-05 (1.6e-06)	3.4e-05	0.0078	0.237	0.275
sensitivity 99%	Ridge	1,000,000	0.0046	stratified	5.5e-05 (0.00029)	3.4e-05 (1.5e-06)	3.4e-05	0.0079	0.235	0.274
sensitivity 99%	Ridge	1,000,000	0.0092	naive	-0.00018 (0.00019)	1.1e-05 (4.9e-07)	1.1e-05	0.0043	0.233	0.254
sensitivity 99%	Ridge	1,000,000	0.0092	stratified	-0.00021 (0.00019)	1.2e-05 (5.1e-07)	1.2e-05	0.0046	0.233	0.254

Table C.21: Bias, MSE, and Variability for sensitivity at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 90%	Logistic	5,000	0.0046	naive	0.16 (0.0089)	0.11 (0.0021)	0.082	0.43	0	1
sensitivity 90%	Logistic	5,000	0.0046	stratified	0.16 (0.009)	0.11 (0.002)	0.083	0.42	0	1
sensitivity 90%	Logistic	5,000	0.0092	naive	-0.0009 (0.0039)	0.025 (0.0021)	0.025	0.13	0.0686	1
sensitivity 90%	Logistic	5,000	0.0092	stratified	0.0015 (0.0037)	0.022 (0.0021)	0.022	0.12	0.065	1
sensitivity 90%	Logistic	5,000	0.0184	naive	-0.012 (0.0024)	0.0068 (0.00082)	0.0067	0.083	0.198	0.967
sensitivity 90%	Logistic	5,000	0.0184	stratified	-0.013 (0.0025)	0.0068 (0.00073)	0.0066	0.079	0.239	0.967
sensitivity 90%	Logistic	10,000	0.0046	naive	-0.017 (0.003)	0.0079 (0.0005)	0.0076	0.12	0.051	0.86
sensitivity 90%	Logistic	10,000	0.0046	stratified	-0.015 (0.0027)	0.0066 (0.00041)	0.0064	0.1	0.08	0.72
sensitivity 90%	Logistic	10,000	0.0092	naive	-0.02 (0.0021)	0.0044 (0.0002)	0.004	0.084	0.216	0.686
sensitivity 90%	Logistic	10,000	0.0092	stratified	-0.021 (0.002)	0.0037 (0.00017)	0.0032	0.077	0.196	0.684
sensitivity 90%	Logistic	10,000	0.0184	naive	-0.013 (0.0014)	0.0018 (7.8e-05)	0.0016	0.056	0.38	0.626
sensitivity 90%	Logistic	10,000	0.0184	stratified	-0.015 (0.0013)	0.0018 (8.6e-05)	0.0016	0.054	0.326	0.609
sensitivity 90%	Logistic	50,000	0.0046	naive	-0.01 (0.0012)	0.0013 (6e-05)	0.0012	0.046	0.428	0.647
sensitivity 90%	Logistic	50,000	0.0046	stratified	-0.01 (0.0012)	0.0013 (6e-05)	0.0012	0.043	0.43	0.665
sensitivity 90%	Logistic	50,000	0.0092	naive	-0.0058 (0.00082)	0.00064 (2.9e-05)	0.00061	0.033	0.522	0.671
sensitivity 90%	Logistic	50,000	0.0092	stratified	-0.0036 (0.00077)	0.00055 (2.5e-05)	0.00054	0.033	0.517	0.674
sensitivity 90%	Logistic	50,000	0.0184	naive	-0.0029 (0.00055)	0.00028 (1.3e-05)	0.00027	0.022	0.55	0.66
sensitivity 90%	Logistic	50,000	0.0184	stratified	-0.0025 (0.00053)	0.00027 (1.2e-05)	0.00027	0.022	0.558	0.665
sensitivity 90%	Logistic	100,000	0.0046	naive	-0.0042 (0.00082)	0.00065 (3e-05)	0.00063	0.033	0.51	0.69
sensitivity 90%	Logistic	100,000	0.0046	stratified	-0.0047 (0.00081)	0.00059 (2.6e-05)	0.00057	0.033	0.524	0.67
sensitivity 90%	Logistic	100,000	0.0092	naive	-0.0022 (0.00054)	0.00027 (1.2e-05)	0.00026	0.021	0.563	0.66
sensitivity 90%	Logistic	100,000	0.0092	stratified	-0.003 (0.00056)	0.00028 (1.3e-05)	0.00027	0.023	0.546	0.664
sensitivity 90%	Logistic	100,000	0.0184	naive	-0.00084 (0.00038)	0.00012 (5.3e-06)	0.00012	0.015	0.588	0.652
sensitivity 90%	Logistic	100,000	0.0184	stratified	-0.0016 (0.00038)	0.00012 (4.9e-06)	0.00011	0.015	0.586	0.649
sensitivity 90%	Logistic	1,000,000	0.0046	naive	-0.00032 (0.00031)	4.3e-05 (1.9e-06)	4.3e-05	0.0093	0.62	0.663
sensitivity 90%	Logistic	1,000,000	0.0046	stratified	-0.00022 (0.00032)	4.1e-05 (1.8e-06)	4.1e-05	0.0091	0.623	0.66
sensitivity 90%	Logistic	1,000,000	0.0092	naive	-9.4e-05 (0.00023)	1.7e-05 (7.8e-07)	1.7e-05	0.0055	0.626	0.653
sensitivity 90%	Logistic	1,000,000	0.0092	stratified	1.6e-05 (0.00022)	1.7e-05 (7.5e-07)	1.7e-05	0.0057	0.624	0.652

Table C.22: Bias, MSE, and Variability for sensitivity at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 95%	Logistic	5,000	0.0046	naive	0.18 (0.0094)	0.12 (0.0023)	0.09	0.47	0	1
sensitivity 95%	Logistic	5,000	0.0046	stratified	0.17 (0.0096)	0.12 (0.0022)	0.093	0.49	0	1
sensitivity 95%	Logistic	5,000	0.0092	naive	0.0074 (0.004)	0.025 (0.0024)	0.025	0.13	0.0143	1
sensitivity 95%	Logistic	5,000	0.0092	stratified	0.0057 (0.0037)	0.022 (0.0024)	0.022	0.11	0	1
sensitivity 95%	Logistic	5,000	0.0184	naive	-0.0066 (0.0025)	0.0072 (0.0011)	0.0072	0.078	0.125	0.967
sensitivity 95%	Logistic	5,000	0.0184	stratified	-0.0089 (0.0025)	0.0072 (0.00096)	0.0071	0.077	0.139	0.944
sensitivity 95%	Logistic	10,000	0.0046	naive	-0.014 (0.0026)	0.0066 (0.00057)	0.0064	0.098	0	0.84
sensitivity 95%	Logistic	10,000	0.0046	stratified	-0.01 (0.0024)	0.0054 (0.0004)	0.0053	0.095	0.045	0.695
sensitivity 95%	Logistic	10,000	0.0092	naive	-0.017 (0.0019)	0.0037 (0.00018)	0.0034	0.079	0.102	0.61
sensitivity 95%	Logistic	10,000	0.0092	stratified	-0.017 (0.0018)	0.0031 (0.00017)	0.0029	0.074	0.151	0.642
sensitivity 95%	Logistic	10,000	0.0184	naive	-0.01 (0.0013)	0.0016 (6.8e-05)	0.0015	0.053	0.261	0.5
sensitivity 95%	Logistic	10,000	0.0184	stratified	-0.013 (0.0012)	0.0015 (6.8e-05)	0.0014	0.05	0.25	0.506
sensitivity 95%	Logistic	50,000	0.0046	naive	-0.0095 (0.0011)	0.0013 (5.7e-05)	0.0012	0.046	0.313	0.517
sensitivity 95%	Logistic	50,000	0.0046	stratified	-0.0099 (0.0011)	0.0013 (5.8e-05)	0.0012	0.043	0.304	0.539
sensitivity 95%	Logistic	50,000	0.0092	naive	-0.0051 (0.0008)	0.00062 (3e-05)	0.00059	0.031	0.38	0.537
sensitivity 95%	Logistic	50,000	0.0092	stratified	-0.0035 (0.00077)	0.00055 (2.5e-05)	0.00054	0.033	0.389	0.539
sensitivity 95%	Logistic	50,000	0.0184	naive	-0.0028 (0.00054)	0.00027 (1.2e-05)	0.00026	0.023	0.418	0.519
sensitivity 95%	Logistic	50,000	0.0184	stratified	-0.002 (0.00055)	0.00028 (1.3e-05)	0.00027	0.022	0.418	0.532
sensitivity 95%	Logistic	100,000	0.0046	naive	-0.004 (0.00084)	0.00066 (2.9e-05)	0.00064	0.034	0.396	0.557
sensitivity 95%	Logistic	100,000	0.0046	stratified	-0.0043 (0.00082)	0.00062 (2.7e-05)	0.0006	0.033	0.396	0.557
sensitivity 95%	Logistic	100,000	0.0092	naive	-0.0023 (0.00058)	0.0003 (1.2e-05)	0.00029	0.024	0.433	0.533
sensitivity 95%	Logistic	100,000	0.0092	stratified	-0.0027 (0.00056)	0.00028 (1.3e-05)	0.00027	0.021	0.431	0.534
sensitivity 95%	Logistic	100,000	0.0184	naive	-0.00058 (0.00039)	0.00013 (5.3e-06)	0.00013	0.016	0.449	0.516
sensitivity 95%	Logistic	100,000	0.0184	stratified	-0.0011 (0.00038)	0.00011 (4.9e-06)	0.00011	0.015	0.445	0.506
sensitivity 95%	Logistic	1,000,000	0.0046	naive	-0.00014 (0.00033)	4.6e-05 (2.1e-06)	4.6e-05	0.0093	0.488	0.534
sensitivity 95%	Logistic	1,000,000	0.0046	stratified	-0.00019 (0.00034)	4.8e-05 (2.1e-06)	4.8e-05	0.0091	0.491	0.53
sensitivity 95%	Logistic	1,000,000	0.0092	naive	7e-05 (0.00023)	1.8e-05 (8.7e-07)	1.8e-05	0.0057	0.492	0.521
sensitivity 95%	Logistic	1,000,000	0.0092	stratified	0.00013 (0.00023)	1.9e-05 (8.3e-07)	1.9e-05	0.0059	0.489	0.52

Table C.23: Bias, MSE, and Variability for sensitivity at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
sensitivity 99%	Logistic	5,000	0.0046	naive	0.033 (0.012)	0.053 (0.0022)	0.052	0.35	0	1
sensitivity 99%	Logistic	5,000	0.0046	stratified	0.023 (0.013)	0.055 (0.0021)	0.054	0.35	0	1
sensitivity 99%	Logistic	5,000	0.0092	naive	0.02 (0.0038)	0.025 (0.0028)	0.025	0.11	0	1
sensitivity 99%	Logistic	5,000	0.0092	stratified	0.018 (0.0036)	0.023 (0.0029)	0.023	0.09	0	1
sensitivity 99%	Logistic	5,000	0.0184	naive	0.0003 (0.0023)	0.0043 (0.00087)	0.0043	0.056	0.0324	0.845
sensitivity 99%	Logistic	5,000	0.0184	stratified	-0.0014 (0.0025)	0.0052 (0.00084)	0.0052	0.048	0.0322	0.753
sensitivity 99%	Logistic	10,000	0.0046	naive	-0.0052 (0.0021)	0.004 (0.00046)	0.004	0.072	0	0.645
sensitivity 99%	Logistic	10,000	0.0046	stratified	-0.003 (0.0018)	0.0035 (0.0005)	0.0035	0.065	0	0.695
sensitivity 99%	Logistic	10,000	0.0092	naive	-0.0061 (0.0014)	0.0018 (0.00013)	0.0018	0.053	0.0341	0.467
sensitivity 99%	Logistic	10,000	0.0092	stratified	-0.0068 (0.0014)	0.0018 (0.00021)	0.0018	0.054	0.01	0.598
sensitivity 99%	Logistic	10,000	0.0184	naive	-0.0041 (0.00094)	0.00078 (3.7e-05)	0.00077	0.036	0.0816	0.28
sensitivity 99%	Logistic	10,000	0.0184	stratified	-0.006 (0.00092)	0.00076 (3.7e-05)	0.00072	0.033	0.081	0.304
sensitivity 99%	Logistic	50,000	0.0046	naive	-0.0048 (0.00091)	0.0008 (3.3e-05)	0.00077	0.04	0.116	0.293
sensitivity 99%	Logistic	50,000	0.0046	stratified	-0.0063 (0.0009)	0.00081 (3.4e-05)	0.00077	0.039	0.117	0.291
sensitivity 99%	Logistic	50,000	0.0092	naive	-0.0035 (0.00065)	0.00038 (1.6e-05)	0.00037	0.026	0.16	0.271
sensitivity 99%	Logistic	50,000	0.0092	stratified	-0.0023 (0.00065)	0.00035 (1.6e-05)	0.00035	0.024	0.159	0.285
sensitivity 99%	Logistic	50,000	0.0184	naive	-0.0016 (0.00045)	0.00014 (6.1e-06)	0.00014	0.017	0.167	0.246
sensitivity 99%	Logistic	50,000	0.0184	stratified	-0.0013 (0.00044)	0.00015 (6.5e-06)	0.00014	0.016	0.165	0.247
sensitivity 99%	Logistic	100,000	0.0046	naive	-0.0026 (0.00069)	0.00042 (1.8e-05)	0.00041	0.028	0.154	0.299
sensitivity 99%	Logistic	100,000	0.0046	stratified	-0.0024 (0.00068)	0.00042 (1.8e-05)	0.00042	0.028	0.167	0.298
sensitivity 99%	Logistic	100,000	0.0092	naive	-0.0016 (0.00048)	0.00019 (8.3e-06)	0.00019	0.018	0.19	0.275
sensitivity 99%	Logistic	100,000	0.0092	stratified	-0.0025 (0.00047)	0.00018 (8.5e-06)	0.00018	0.017	0.184	0.275
sensitivity 99%	Logistic	100,000	0.0184	naive	-5.3e-05 (0.00031)	6.7e-05 (3e-06)	6.7e-05	0.011	0.181	0.242
sensitivity 99%	Logistic	100,000	0.0184	stratified	-0.00088 (0.00031)	6.5e-05 (3e-06)	6.4e-05	0.01	0.187	0.236
sensitivity 99%	Logistic	1,000,000	0.0046	naive	-0.00023 (0.00029)	3.5e-05 (1.5e-06)	3.5e-05	0.0078	0.238	0.274
sensitivity 99%	Logistic	1,000,000	0.0046	stratified	-0.00063 (0.00029)	3.4e-05 (1.5e-06)	3.4e-05	0.0078	0.24	0.277
sensitivity 99%	Logistic	1,000,000	0.0092	naive	0.00014 (0.0002)	1.3e-05 (6e-07)	1.3e-05	0.005	0.234	0.258
sensitivity 99%	Logistic	1,000,000	0.0092	stratified	0.00011 (0.0002)	1.3e-05 (5.6e-07)	1.3e-05	0.0047	0.232	0.253

Table C.24: Bias, MSE, and Variability for sensitivity at the 99th percentile, logistic regression.

C.5 Specificity

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 90%	Random Forest	5,000	0.0046	naive	-0.00024 (0.00022)	8.4e-06 (6e-07)	8.4e-06	0.0037	0.887	0.907
specificity 90%	Random Forest	5,000	0.0046	stratified	4.2e-05 (0.00022)	5.6e-06 (3.7e-07)	5.6e-06	0.003	0.889	0.904
specificity 90%	Random Forest	5,000	0.0092	naive	-0.00012 (0.00021)	4.7e-06 (2.9e-07)	4.7e-06	0.0028	0.892	0.906
specificity 90%	Random Forest	5,000	0.0092	stratified	-0.00022 (0.00022)	4.6e-06 (2.9e-07)	4.5e-06	0.003	0.893	0.906
specificity 90%	Random Forest	5,000	0.0184	naive	-8.3e-05 (0.00021)	4e-06 (2.7e-07)	4e-06	0.0026	0.899	0.911
specificity 90%	Random Forest	5,000	0.0184	stratified	0.00014 (0.0002)	3.6e-06 (2.3e-07)	3.6e-06	0.0024	0.899	0.911
specificity 90%	Random Forest	10,000	0.0046	naive	-0.00031 (0.00014)	1.8e-06 (1.2e-07)	1.7e-06	0.0017	0.897	0.905
specificity 90%	Random Forest	10,000	0.0046	stratified	-0.00019 (0.00015)	1.7e-06 (1.1e-07)	1.6e-06	0.0016	0.898	0.905
specificity 90%	Random Forest	10,000	0.0092	naive	-3.3e-05 (0.00014)	2.2e-06 (1.4e-07)	2.2e-06	0.002	0.894	0.904
specificity 90%	Random Forest	10,000	0.0092	stratified	-0.00013 (0.00015)	2.1e-06 (1.3e-07)	2.1e-06	0.0019	0.895	0.904
specificity 90%	Random Forest	10,000	0.0184	naive	-9.8e-05 (0.00014)	1.9e-06 (1.2e-07)	1.9e-06	0.0018	0.901	0.91
specificity 90%	Random Forest	10,000	0.0184	stratified	0.00013 (0.00014)	1.9e-06 (1.2e-07)	1.9e-06	0.0019	0.901	0.91
specificity 90%	Random Forest	50,000	0.0046	naive	-4.7e-05 (6.5e-05)	2.4e-07 (1.5e-08)	2.4e-07	0.00068	0.9	0.903
specificity 90%	Random Forest	50,000	0.0046	stratified	-5.3e-05 (6.4e-05)	2.2e-07 (1.4e-08)	2.2e-07	0.0006	0.9	0.903
specificity 90%	Random Forest	50,000	0.0092	naive	5e-05 (6.5e-05)	2.1e-07 (1.3e-08)	2.1e-07	0.00061	0.903	0.905
specificity 90%	Random Forest	50,000	0.0092	stratified	2.8e-05 (6.3e-05)	2.1e-07 (1.3e-08)	2.1e-07	0.00063	0.903	0.906
specificity 90%	Random Forest	50,000	0.0184	naive	3.5e-05 (6.1e-05)	2.2e-07 (1.4e-08)	2.2e-07	0.00069	0.907	0.91
specificity 90%	Random Forest	50,000	0.0184	stratified	-8e-05 (5.9e-05)	2.2e-07 (1.3e-08)	2.2e-07	0.00061	0.908	0.91
specificity 90%	Random Forest	100,000	0.0046	naive	2.4e-05 (4.9e-05)	1.4e-07 (8.9e-09)	1.4e-07	0.0005	0.901	0.903
specificity 90%	Random Forest	100,000	0.0046	stratified	-5e-05 (4.5e-05)	1.4e-07 (8.2e-09)	1.4e-07	0.00056	0.901	0.903
specificity 90%	Random Forest	100,000	0.0092	naive	6.9e-05 (4.7e-05)	1.3e-07 (8.5e-09)	1.3e-07	0.00044	0.903	0.905
specificity 90%	Random Forest	100,000	0.0092	stratified	-8.8e-06 (4.6e-05)	1.1e-07 (6.9e-09)	1.1e-07	0.00044	0.903	0.905
specificity 90%	Random Forest	100,000	0.0184	naive	-4.2e-05 (4.3e-05)	1.2e-07 (8.4e-09)	1.1e-07	0.00043	0.908	0.91
specificity 90%	Random Forest	100,000	0.0184	stratified	-1.2e-05 (4.5e-05)	1.1e-07 (7.3e-09)	1.1e-07	0.00046	0.908	0.91
specificity 90%	Random Forest	1,000,000	0.0046	naive	-6.6e-06 (2e-05)	1.1e-08 (8.5e-10)	1.1e-08	0.00014	0.902	0.903
specificity 90%	Random Forest	1,000,000	0.0046	stratified	-2.6e-05 (2e-05)	1e-08 (6.3e-10)	9.4e-09	0.00013	0.902	0.903
specificity 90%	Random Forest	1,000,000	0.0092	naive	-2.7e-05 (1.9e-05)	1.1e-08 (7e-10)	9.9e-09	0.00013	0.905	0.905
specificity 90%	Random Forest	1,000,000	0.0092	stratified	-9.9e-06 (2e-05)	9.1e-09 (5.7e-10)	9e-09	0.00012	0.905	0.905

Table C.25: Bias, MSE, and Variability for specificity at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 95%	Random Forest	5,000	0.0046	naive	-0.00023 (0.00016)	4.5e-06 (3.1e-07)	4.5e-06	0.0027	0.941	0.956
specificity 95%	Random Forest	5,000	0.0046	stratified	-0.0002 (0.00016)	2.7e-06 (1.8e-07)	2.7e-06	0.0022	0.942	0.953
specificity 95%	Random Forest	5,000	0.0092	naive	-7.6e-05 (0.00015)	2.4e-06 (1.6e-07)	2.4e-06	0.002	0.946	0.956
specificity 95%	Random Forest	5,000	0.0092	stratified	-0.00035 (0.00015)	2.4e-06 (1.5e-07)	2.3e-06	0.002	0.947	0.955
specificity 95%	Random Forest	5,000	0.0184	naive	-2e-05 (0.00015)	2.5e-06 (1.6e-07)	2.5e-06	0.0022	0.952	0.961
specificity 95%	Random Forest	5,000	0.0184	stratified	5.8e-05 (0.00014)	2.4e-06 (1.5e-07)	2.4e-06	0.002	0.951	0.96
specificity 95%	Random Forest	10,000	0.0046	naive	-3.8e-08 (0.00011)	1.1e-06 (7e-08)	1.1e-06	0.0014	0.948	0.956
specificity 95%	Random Forest	10,000	0.0046	stratified	-0.00014 (0.00011)	1.1e-06 (7.4e-08)	1e-06	0.0013	0.947	0.954
specificity 95%	Random Forest	10,000	0.0092	naive	-2.9e-05 (0.0001)	1.1e-06 (6.9e-08)	1.1e-06	0.0015	0.949	0.955
specificity 95%	Random Forest	10,000	0.0092	stratified	2.1e-06 (9.9e-05)	1.1e-06 (6.6e-08)	1.1e-06	0.0015	0.949	0.955
specificity 95%	Random Forest	10,000	0.0184	naive	-7.6e-06 (9.3e-05)	9.4e-07 (5.7e-08)	9.4e-07	0.0014	0.954	0.96
specificity 95%	Random Forest	10,000	0.0184	stratified	-0.00012 (9.9e-05)	9.9e-07 (6e-08)	9.8e-07	0.0014	0.954	0.959
specificity 95%	Random Forest	50,000	0.0046	naive	2.1e-05 (4.8e-05)	1.5e-07 (9.7e-09)	1.5e-07	0.00052	0.95	0.953
specificity 95%	Random Forest	50,000	0.0046	stratified	-2.9e-05 (4.5e-05)	1.4e-07 (9.4e-09)	1.4e-07	0.00048	0.95	0.953
specificity 95%	Random Forest	50,000	0.0092	naive	-7e-05 (4.6e-05)	1.4e-07 (9.1e-09)	1.3e-07	0.00052	0.952	0.955
specificity 95%	Random Forest	50,000	0.0092	stratified	-3.2e-05 (4.6e-05)	1.3e-07 (8.4e-09)	1.3e-07	0.00052	0.952	0.955
specificity 95%	Random Forest	50,000	0.0184	naive	-2.4e-05 (4.4e-05)	1.5e-07 (8.9e-09)	1.5e-07	0.00051	0.957	0.959
specificity 95%	Random Forest	50,000	0.0184	stratified	-8.4e-06 (4.3e-05)	1.5e-07 (9.3e-09)	1.5e-07	0.00053	0.957	0.959
specificity 95%	Random Forest	100,000	0.0046	naive	-7.5e-06 (3.3e-05)	7.4e-08 (4.8e-09)	7.4e-08	0.00033	0.951	0.953
specificity 95%	Random Forest	100,000	0.0046	stratified	-6.8e-05 (3.1e-05)	9.3e-08 (5.6e-09)	8.8e-08	0.00038	0.951	0.952
specificity 95%	Random Forest	100,000	0.0092	naive	3.5e-05 (3.2e-05)	8.2e-08 (5.1e-09)	8.1e-08	0.00039	0.953	0.955
specificity 95%	Random Forest	100,000	0.0092	stratified	2e-06 (3.3e-05)	6.8e-08 (4.2e-09)	6.9e-08	0.00034	0.953	0.955
specificity 95%	Random Forest	100,000	0.0184	naive	-2.6e-05 (3.2e-05)	8.8e-08 (5.4e-09)	8.7e-08	0.00041	0.957	0.959
specificity 95%	Random Forest	100,000	0.0184	stratified	3e-05 (3.1e-05)	8.7e-08 (5.4e-09)	8.6e-08	0.0004	0.957	0.959
specificity 95%	Random Forest	1,000,000	0.0046	naive	-1.1e-05 (1.4e-05)	6.2e-09 (3.6e-10)	6.1e-09	0.00011	0.952	0.952
specificity 95%	Random Forest	1,000,000	0.0046	stratified	-3.4e-06 (1.4e-05)	6.1e-09 (3.9e-10)	6.1e-09	0.0001	0.952	0.952
specificity 95%	Random Forest	1,000,000	0.0092	naive	-1.1e-05 (1.3e-05)	6.1e-09 (3.7e-10)	6e-09	0.00011	0.954	0.954
specificity 95%	Random Forest	1,000,000	0.0092	stratified	-3.8e-05 (1.4e-05)	7.5e-09 (4.9e-10)	6e-09	0.00011	0.954	0.954

Table C.26: Bias, MSE, and Variability for specificity at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 99%	Random Forest	5,000	0.0046	naive	-0.00019 (7.6e-05)	1e-06 (8e-08)	9.7e-07	0.0013	0.987	0.994
specificity 99%	Random Forest	5,000	0.0046	stratified	5.5e-05 (6.6e-05)	6.5e-07 (3.9e-08)	6.4e-07	0.001	0.989	0.993
specificity 99%	Random Forest	5,000	0.0092	naive	-5e-05 (6.1e-05)	7.2e-07 (4.4e-08)	7.1e-07	0.0012	0.99	0.995
specificity 99%	Random Forest	5,000	0.0092	stratified	-9.8e-05 (5.9e-05)	6.5e-07 (4.8e-08)	6.5e-07	0.001	0.99	0.995
specificity 99%	Random Forest	5,000	0.0184	naive	-2.2e-05 (4.8e-05)	5.5e-07 (3.9e-08)	5.5e-07	0.001	0.992	0.997
specificity 99%	Random Forest	5,000	0.0184	stratified	-4.2e-06 (4.7e-05)	5.2e-07 (3.1e-08)	5.2e-07	0.001	0.992	0.996
specificity 99%	Random Forest	10,000	0.0046	naive	-8.5e-06 (5.1e-05)	3e-07 (2.1e-08)	3e-07	0.00071	0.989	0.992
specificity 99%	Random Forest	10,000	0.0046	stratified	-0.00012 (4.7e-05)	2.9e-07 (2e-08)	2.8e-07	0.0008	0.989	0.992
specificity 99%	Random Forest	10,000	0.0092	naive	-8.8e-05 (4.3e-05)	3e-07 (2e-08)	3e-07	0.0008	0.991	0.994
specificity 99%	Random Forest	10,000	0.0092	stratified	-1.8e-05 (4.2e-05)	2.7e-07 (1.8e-08)	2.7e-07	0.00071	0.991	0.994
specificity 99%	Random Forest	10,000	0.0184	naive	-3.4e-05 (3.4e-05)	2.2e-07 (1.4e-08)	2.2e-07	0.00061	0.993	0.996
specificity 99%	Random Forest	10,000	0.0184	stratified	1.7e-05 (3.5e-05)	2.4e-07 (1.5e-08)	2.4e-07	0.00071	0.993	0.996
specificity 99%	Random Forest	50,000	0.0046	naive	-5.9e-06 (2e-05)	3.9e-08 (2.4e-09)	3.9e-08	0.00028	0.99	0.992
specificity 99%	Random Forest	50,000	0.0046	stratified	-1.5e-05 (2e-05)	3.5e-08 (2.5e-09)	3.5e-08	0.00024	0.99	0.992
specificity 99%	Random Forest	50,000	0.0092	naive	-1e-05 (1.9e-05)	3.6e-08 (2.2e-09)	3.6e-08	0.00024	0.992	0.993
specificity 99%	Random Forest	50,000	0.0092	stratified	-1.2e-05 (1.8e-05)	3.8e-08 (2.5e-09)	3.8e-08	0.00026	0.991	0.993
specificity 99%	Random Forest	50,000	0.0184	naive	8.9e-06 (1.6e-05)	4e-08 (2.6e-09)	4e-08	0.00029	0.993	0.994
specificity 99%	Random Forest	50,000	0.0184	stratified	-4.9e-06 (1.7e-05)	4.3e-08 (2.8e-09)	4.3e-08	0.00029	0.993	0.994
specificity 99%	Random Forest	100,000	0.0046	naive	-2.1e-05 (1.4e-05)	2.2e-08 (1.3e-09)	2.2e-08	0.00021	0.991	0.992
specificity 99%	Random Forest	100,000	0.0046	stratified	-2.9e-05 (1.4e-05)	2.2e-08 (1.5e-09)	2.1e-08	0.00018	0.991	0.992
specificity 99%	Random Forest	100,000	0.0092	naive	6.2e-06 (1.3e-05)	2e-08 (1.3e-09)	2e-08	0.00019	0.992	0.993
specificity 99%	Random Forest	100,000	0.0092	stratified	-1.2e-05 (1.4e-05)	1.9e-08 (1.3e-09)	1.9e-08	0.00017	0.992	0.993
specificity 99%	Random Forest	100,000	0.0184	naive	-1.4e-05 (1.3e-05)	2.1e-08 (1.4e-09)	2.1e-08	0.00019	0.993	0.994
specificity 99%	Random Forest	100,000	0.0184	stratified	1.4e-07 (1.2e-05)	2.2e-08 (1.3e-09)	2.2e-08	0.0002	0.993	0.994
specificity 99%	Random Forest	1,000,000	0.0046	naive	1.6e-06 (6.5e-06)	1.7e-09 (1.1e-10)	1.7e-09	5.4e-05	0.991	0.991
specificity 99%	Random Forest	1,000,000	0.0046	stratified	6.1e-06 (6.1e-06)	1.3e-09 (9.1e-11)	1.3e-09	4.5e-05	0.991	0.991
specificity 99%	Random Forest	1,000,000	0.0092	naive	-2e-06 (5.9e-06)	1.6e-09 (9.8e-11)	1.6e-09	5.5e-05	0.992	0.992
specificity 99%	Random Forest	1,000,000	0.0092	stratified	-4.4e-06 (5.8e-06)	1.9e-09 (1.2e-10)	1.8e-09	6.2e-05	0.992	0.992

Table C.27: Bias, MSE, and Variability for specificity at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 90%	Ridge	5,000	0.0046	naive	-7.1e-05 (0.00015)	4.3e-06 (2.5e-07)	4.3e-06	0.0026	0.891	0.909
specificity 90%	Ridge	5,000	0.0046	stratified	-1.2e-05 (0.00014)	2.2e-06 (1.1e-07)	2.2e-06	0.002	0.895	0.907
specificity 90%	Ridge	5,000	0.0092	naive	-0.00014 (0.00014)	2.5e-06 (1.3e-07)	2.5e-06	0.0022	0.898	0.91
specificity 90%	Ridge	5,000	0.0092	stratified	-0.00021 (0.00014)	2.2e-06 (1e-07)	2.2e-06	0.0019	0.899	0.909
specificity 90%	Ridge	5,000	0.0184	naive	-0.00016 (0.00013)	2.6e-06 (1.2e-07)	2.6e-06	0.0021	0.903	0.913
specificity 90%	Ridge	5,000	0.0184	stratified	-0.00012 (0.00014)	2.5e-06 (1.2e-07)	2.5e-06	0.002	0.903	0.914
specificity 90%	Ridge	10,000	0.0046	naive	-1.1e-05 (9.9e-05)	9.7e-07 (4.5e-08)	9.8e-07	0.0013	0.899	0.905
specificity 90%	Ridge	10,000	0.0046	stratified	-0.00014 (0.0001)	1.1e-06 (4.7e-08)	1.1e-06	0.0013	0.899	0.905
specificity 90%	Ridge	10,000	0.0092	naive	-6.3e-06 (9.7e-05)	1.1e-06 (4.7e-08)	1.1e-06	0.0014	0.901	0.907
specificity 90%	Ridge	10,000	0.0092	stratified	-3.7e-05 (9.7e-05)	1.2e-06 (5.4e-08)	1.2e-06	0.0015	0.901	0.908
specificity 90%	Ridge	10,000	0.0184	naive	6.3e-05 (9.5e-05)	1.4e-06 (6.4e-08)	1.4e-06	0.0015	0.905	0.913
specificity 90%	Ridge	10,000	0.0184	stratified	-0.00017 (9.7e-05)	1.3e-06 (5.9e-08)	1.3e-06	0.0015	0.905	0.913
specificity 90%	Ridge	50,000	0.0046	naive	-3.4e-05 (4.6e-05)	2e-07 (8.6e-09)	2e-07	0.00062	0.901	0.904
specificity 90%	Ridge	50,000	0.0046	stratified	-1.6e-05 (4.4e-05)	2.2e-07 (8.9e-09)	2.2e-07	0.00064	0.901	0.904
specificity 90%	Ridge	50,000	0.0092	naive	-2.6e-05 (4.5e-05)	2.1e-07 (8.8e-09)	2.1e-07	0.00062	0.903	0.906
specificity 90%	Ridge	50,000	0.0092	stratified	1.1e-06 (4.5e-05)	2.2e-07 (9.8e-09)	2.2e-07	0.00061	0.903	0.906
specificity 90%	Ridge	50,000	0.0184	naive	-2.7e-05 (4.5e-05)	2.3e-07 (1e-08)	2.3e-07	0.00068	0.908	0.912
specificity 90%	Ridge	50,000	0.0184	stratified	-2.1e-05 (4.4e-05)	2.4e-07 (1e-08)	2.4e-07	0.00065	0.908	0.911
specificity 90%	Ridge	100,000	0.0046	naive	7.2e-06 (3.3e-05)	9e-08 (3.9e-09)	9e-08	0.00042	0.902	0.903
specificity 90%	Ridge	100,000	0.0046	stratified	4e-05 (3.2e-05)	8.9e-08 (4e-09)	8.8e-08	0.00039	0.901	0.903
specificity 90%	Ridge	100,000	0.0092	naive	4.2e-05 (3.2e-05)	9.6e-08 (4.4e-09)	9.5e-08	0.00042	0.904	0.906
specificity 90%	Ridge	100,000	0.0092	stratified	-2.3e-09 (3.3e-05)	9.5e-08 (4.2e-09)	9.5e-08	0.00039	0.904	0.906
specificity 90%	Ridge	100,000	0.0184	naive	3.9e-05 (3.2e-05)	9.3e-08 (4.1e-09)	9.1e-08	0.00041	0.909	0.911
specificity 90%	Ridge	100,000	0.0184	stratified	5e-05 (3.2e-05)	9.2e-08 (4.2e-09)	9e-08	0.00039	0.909	0.911
specificity 90%	Ridge	1,000,000	0.0046	naive	-6.2e-06 (1.4e-05)	4e-09 (1.7e-10)	4e-09	8.7e-05	0.902	0.903
specificity 90%	Ridge	1,000,000	0.0046	stratified	1.4e-05 (1.4e-05)	4.4e-09 (1.8e-10)	4.2e-09	9.1e-05	0.902	0.903
specificity 90%	Ridge	1,000,000	0.0092	naive	-3.4e-05 (1.4e-05)	4.8e-09 (2.2e-10)	3.6e-09	7.7e-05	0.905	0.905
specificity 90%	Ridge	1,000,000	0.0092	stratified	-3.5e-05 (1.3e-05)	5.2e-09 (2.2e-10)	3.9e-09	8.1e-05	0.905	0.905

Table C.28: Bias, MSE, and Variability for specificity at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 95%	Ridge	5,000	0.0046	naive	-0.00019 (0.00011)	2.9e-06 (1.4e-07)	2.9e-06	0.0022	0.945	0.957
specificity 95%	Ridge	5,000	0.0046	stratified	-7.9e-05 (0.0001)	1.4e-06 (6.2e-08)	1.4e-06	0.0016	0.947	0.955
specificity 95%	Ridge	5,000	0.0092	naive	-0.00012 (9.8e-05)	1.6e-06 (7.2e-08)	1.6e-06	0.0018	0.949	0.957
specificity 95%	Ridge	5,000	0.0092	stratified	-0.0001 (1e-04)	1.5e-06 (6.5e-08)	1.5e-06	0.0016	0.95	0.957
specificity 95%	Ridge	5,000	0.0184	naive	-4.7e-05 (9.4e-05)	1.9e-06 (8.3e-08)	1.9e-06	0.002	0.953	0.962
specificity 95%	Ridge	5,000	0.0184	stratified	6.5e-05 (9.9e-05)	1.8e-06 (7.7e-08)	1.8e-06	0.0018	0.953	0.961
specificity 95%	Ridge	10,000	0.0046	naive	-1e-06 (7.3e-05)	6.6e-07 (3e-08)	6.6e-07	0.0011	0.949	0.955
specificity 95%	Ridge	10,000	0.0046	stratified	4.6e-06 (6.9e-05)	5.7e-07 (2.5e-08)	5.7e-07	0.001	0.949	0.954
specificity 95%	Ridge	10,000	0.0092	naive	4.1e-05 (7.1e-05)	6.8e-07 (3.2e-08)	6.8e-07	0.0011	0.951	0.956
specificity 95%	Ridge	10,000	0.0092	stratified	-6.3e-05 (7.2e-05)	7.8e-07 (3.5e-08)	7.8e-07	0.0012	0.95	0.957
specificity 95%	Ridge	10,000	0.0184	naive	0.00011 (7e-05)	8.9e-07 (3.8e-08)	8.8e-07	0.0012	0.955	0.96
specificity 95%	Ridge	10,000	0.0184	stratified	-0.00011 (6.7e-05)	8.6e-07 (3.9e-08)	8.5e-07	0.0012	0.954	0.96
specificity 95%	Ridge	50,000	0.0046	naive	-1.7e-05 (3.3e-05)	1.2e-07 (5.1e-09)	1.2e-07	0.0005	0.951	0.953
specificity 95%	Ridge	50,000	0.0046	stratified	1.5e-05 (3.3e-05)	1.2e-07 (5.6e-09)	1.2e-07	0.00047	0.951	0.953
specificity 95%	Ridge	50,000	0.0092	naive	-1.5e-05 (3.2e-05)	1.2e-07 (5.3e-09)	1.2e-07	0.00048	0.953	0.955
specificity 95%	Ridge	50,000	0.0092	stratified	-2.5e-05 (3.1e-05)	1.2e-07 (6.1e-09)	1.2e-07	0.00044	0.953	0.955
specificity 95%	Ridge	50,000	0.0184	naive	-2.7e-05 (3e-05)	1.5e-07 (7.1e-09)	1.5e-07	0.00051	0.957	0.959
specificity 95%	Ridge	50,000	0.0184	stratified	-1.4e-05 (3.1e-05)	1.5e-07 (7.2e-09)	1.5e-07	0.00051	0.957	0.959
specificity 95%	Ridge	100,000	0.0046	naive	4.6e-06 (2.3e-05)	5.6e-08 (2.5e-09)	5.6e-08	0.00032	0.951	0.953
specificity 95%	Ridge	100,000	0.0046	stratified	1.7e-05 (2.3e-05)	5.4e-08 (2.4e-09)	5.4e-08	0.00032	0.951	0.953
specificity 95%	Ridge	100,000	0.0092	naive	2.5e-05 (2.3e-05)	6e-08 (2.7e-09)	5.9e-08	0.00032	0.953	0.955
specificity 95%	Ridge	100,000	0.0092	stratified	-1.2e-05 (2.3e-05)	5.8e-08 (2.5e-09)	5.8e-08	0.00033	0.953	0.955
specificity 95%	Ridge	100,000	0.0184	naive	2.4e-05 (2.2e-05)	7.6e-08 (3.3e-09)	7.5e-08	0.00038	0.957	0.959
specificity 95%	Ridge	100,000	0.0184	stratified	1.7e-05 (2.2e-05)	6.7e-08 (3e-09)	6.7e-08	0.00035	0.957	0.959
specificity 95%	Ridge	1,000,000	0.0046	naive	-6.8e-06 (9.9e-06)	2.7e-09 (1.2e-10)	2.7e-09	6.7e-05	0.952	0.952
specificity 95%	Ridge	1,000,000	0.0046	stratified	-1.5e-05 (9.5e-06)	3e-09 (1.3e-10)	2.8e-09	7.1e-05	0.952	0.952
specificity 95%	Ridge	1,000,000	0.0092	naive	-5.9e-06 (9.3e-06)	2.9e-09 (1.3e-10)	2.8e-09	7.3e-05	0.954	0.954
specificity 95%	Ridge	1,000,000	0.0092	stratified	-1.2e-05 (9.2e-06)	2.7e-09 (1.1e-10)	2.5e-09	7e-05	0.954	0.954

Table C.29: Bias, MSE, and Variability for specificity at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 99%	Ridge	5,000	0.0046	naive	-6e-05 (4.7e-05)	6.5e-07 (3.1e-08)	6.4e-07	0.001	0.988	0.993
specificity 99%	Ridge	5,000	0.0046	stratified	-7.4e-05 (4.5e-05)	4.8e-07 (2.1e-08)	4.8e-07	0.001	0.989	0.993
specificity 99%	Ridge	5,000	0.0092	naive	-6.7e-05 (4.3e-05)	5.3e-07 (2.5e-08)	5.3e-07	0.001	0.989	0.994
specificity 99%	Ridge	5,000	0.0092	stratified	4.4e-06 (4.3e-05)	5.2e-07 (2.4e-08)	5.2e-07	0.001	0.99	0.994
specificity 99%	Ridge	5,000	0.0184	naive	-5.2e-05 (3.7e-05)	6.1e-07 (2.7e-08)	6.1e-07	0.001	0.991	0.996
specificity 99%	Ridge	5,000	0.0184	stratified	-1.5e-05 (3.8e-05)	6.5e-07 (3e-08)	6.5e-07	0.0012	0.991	0.996
specificity 99%	Ridge	10,000	0.0046	naive	9.3e-06 (3.2e-05)	2e-07 (9.2e-09)	2e-07	0.0006	0.989	0.992
specificity 99%	Ridge	10,000	0.0046	stratified	3.1e-05 (3.3e-05)	2.2e-07 (1e-08)	2.2e-07	0.0006	0.989	0.992
specificity 99%	Ridge	10,000	0.0092	naive	1.8e-05 (3e-05)	2.7e-07 (1.2e-08)	2.7e-07	0.00071	0.99	0.994
specificity 99%	Ridge	10,000	0.0092	stratified	-7.3e-05 (3.1e-05)	2.8e-07 (1.4e-08)	2.7e-07	0.00061	0.99	0.994
specificity 99%	Ridge	10,000	0.0184	naive	4.6e-05 (2.6e-05)	3.8e-07 (1.7e-08)	3.7e-07	0.00081	0.992	0.996
specificity 99%	Ridge	10,000	0.0184	stratified	-1.4e-05 (2.7e-05)	3.4e-07 (1.5e-08)	3.4e-07	0.00081	0.992	0.995
specificity 99%	Ridge	50,000	0.0046	naive	-3.1e-06 (1.4e-05)	3.4e-08 (1.5e-09)	3.4e-08	0.00024	0.99	0.992
specificity 99%	Ridge	50,000	0.0046	stratified	7.5e-06 (1.5e-05)	3.8e-08 (1.6e-09)	3.8e-08	0.00026	0.99	0.992
specificity 99%	Ridge	50,000	0.0092	naive	-1.5e-05 (1.4e-05)	4.7e-08 (2e-09)	4.7e-08	0.00028	0.991	0.993
specificity 99%	Ridge	50,000	0.0092	stratified	-6.7e-06 (1.3e-05)	4.5e-08 (2e-09)	4.5e-08	0.00028	0.991	0.993
specificity 99%	Ridge	50,000	0.0184	naive	1.9e-05 (1.2e-05)	5.6e-08 (2.6e-09)	5.6e-08	0.00031	0.993	0.995
specificity 99%	Ridge	50,000	0.0184	stratified	-2.5e-05 (1.2e-05)	5.6e-08 (2.6e-09)	5.6e-08	0.00031	0.993	0.994
specificity 99%	Ridge	100,000	0.0046	naive	-5.6e-06 (9.9e-06)	1.8e-08 (8.1e-10)	1.8e-08	0.00018	0.991	0.992
specificity 99%	Ridge	100,000	0.0046	stratified	7.4e-06 (1e-05)	1.7e-08 (8.2e-10)	1.7e-08	0.00018	0.991	0.992
specificity 99%	Ridge	100,000	0.0092	naive	-1.1e-06 (9.8e-06)	2.4e-08 (1e-09)	2.4e-08	0.0002	0.992	0.993
specificity 99%	Ridge	100,000	0.0092	stratified	2.7e-06 (9.7e-06)	2.2e-08 (1e-09)	2.2e-08	0.0002	0.992	0.993
specificity 99%	Ridge	100,000	0.0184	naive	-1.9e-05 (8.7e-06)	2.8e-08 (1.3e-09)	2.8e-08	0.00022	0.993	0.994
specificity 99%	Ridge	100,000	0.0184	stratified	1.6e-05 (8.8e-06)	2.7e-08 (1.2e-09)	2.7e-08	0.00021	0.993	0.994
specificity 99%	Ridge	1,000,000	0.0046	naive	4.5e-06 (4.3e-06)	1.1e-09 (5.2e-11)	1.1e-09	4.1e-05	0.991	0.991
specificity 99%	Ridge	1,000,000	0.0046	stratified	-9.5e-06 (4.2e-06)	1.2e-09 (5.5e-11)	1.2e-09	4.6e-05	0.991	0.991
specificity 99%	Ridge	1,000,000	0.0092	naive	-5.5e-06 (3.9e-06)	1.2e-09 (5.5e-11)	1.2e-09	4.4e-05	0.992	0.992
specificity 99%	Ridge	1,000,000	0.0092	stratified	-1e-05 (3.9e-06)	1.3e-09 (6.2e-11)	1.2e-09	4.4e-05	0.992	0.992

Table C.30: Bias, MSE, and Variability for specificity at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 95%	Logistic	5,000	0.0046	naive	-0.19 (0.01)	0.13 (0.0026)	0.097	0.51	0	0.956
specificity 95%	Logistic	5,000	0.0046	stratified	-0.18 (0.01)	0.14 (0.0025)	0.1	0.51	0	0.955
specificity 95%	Logistic	5,000	0.0092	naive	-0.029 (0.0037)	0.029 (0.0036)	0.029	0.095	0	0.959
specificity 95%	Logistic	5,000	0.0092	stratified	-0.023 (0.0033)	0.026 (0.0035)	0.026	0.094	0	0.96
specificity 95%	Logistic	5,000	0.0184	naive	-0.0073 (0.0019)	0.0082 (0.0022)	0.0082	0.0031	0.0213	0.961
specificity 95%	Logistic	5,000	0.0184	stratified	-0.008 (0.0021)	0.0089 (0.0019)	0.0088	0.0029	0.0943	0.963
specificity 95%	Logistic	10,000	0.0046	naive	-0.0047 (0.0013)	0.002 (0.00084)	0.002	0.0021	0.137	0.957
specificity 95%	Logistic	10,000	0.0046	stratified	-0.0032 (0.0012)	0.0017 (0.00057)	0.0017	0.0022	0.367	0.957
specificity 95%	Logistic	10,000	0.0092	naive	-0.00053 (0.00066)	0.00025 (0.00015)	0.00025	0.0018	0.57	0.958
specificity 95%	Logistic	10,000	0.0092	stratified	-0.0012 (0.00054)	0.00029 (0.00025)	0.00029	0.0018	0.457	0.958
specificity 95%	Logistic	10,000	0.0184	naive	-0.00034 (0.00012)	1.1e-05 (8.9e-06)	1.1e-05	0.0017	0.862	0.961
specificity 95%	Logistic	10,000	0.0184	stratified	-0.00038 (0.00016)	2.2e-05 (2e-05)	2.2e-05	0.0017	0.816	0.961
specificity 95%	Logistic	50,000	0.0046	naive	2.5e-05 (3.3e-05)	2e-07 (8.8e-09)	2e-07	0.00061	0.95	0.953
specificity 95%	Logistic	50,000	0.0046	stratified	-2.9e-05 (3.3e-05)	2e-07 (9.3e-09)	2e-07	0.0006	0.95	0.953
specificity 95%	Logistic	50,000	0.0092	naive	-1.7e-05 (3.3e-05)	1.8e-07 (8.2e-09)	1.8e-07	0.00056	0.952	0.955
specificity 95%	Logistic	50,000	0.0092	stratified	-5.2e-05 (3.2e-05)	1.8e-07 (8.1e-09)	1.8e-07	0.00059	0.953	0.955
specificity 95%	Logistic	50,000	0.0184	naive	-1.5e-05 (3e-05)	1.8e-07 (7.7e-09)	1.8e-07	0.00057	0.957	0.959
specificity 95%	Logistic	50,000	0.0184	stratified	-4.8e-05 (3.1e-05)	1.8e-07 (8.3e-09)	1.8e-07	0.00058	0.956	0.959
specificity 95%	Logistic	100,000	0.0046	naive	-2.8e-05 (2.4e-05)	8.4e-08 (3.5e-09)	8.3e-08	0.00039	0.951	0.953
specificity 95%	Logistic	100,000	0.0046	stratified	-3.2e-05 (2.4e-05)	8.1e-08 (3.6e-09)	8.1e-08	0.00037	0.951	0.953
specificity 95%	Logistic	100,000	0.0092	naive	-3.8e-05 (2.4e-05)	7.4e-08 (3.4e-09)	7.3e-08	0.00035	0.953	0.955
specificity 95%	Logistic	100,000	0.0092	stratified	-4.6e-06 (2.4e-05)	7.9e-08 (3.4e-09)	7.9e-08	0.0004	0.953	0.955
specificity 95%	Logistic	100,000	0.0184	naive	-1.1e-05 (2.2e-05)	7.7e-08 (3.2e-09)	7.7e-08	0.00037	0.957	0.959
specificity 95%	Logistic	100,000	0.0184	stratified	-3.8e-05 (2.2e-05)	7.4e-08 (3.1e-09)	7.2e-08	0.00037	0.957	0.959
specificity 95%	Logistic	1,000,000	0.0046	naive	-8.5e-06 (1e-05)	3.3e-09 (1.5e-10)	3.2e-09	7.8e-05	0.952	0.952
specificity 95%	Logistic	1,000,000	0.0046	stratified	1.1e-05 (1e-05)	3.4e-09 (1.4e-10)	3.3e-09	8e-05	0.952	0.952
specificity 95%	Logistic	1,000,000	0.0092	naive	-9.3e-07 (1e-05)	3e-09 (1.4e-10)	3e-09	7.4e-05	0.954	0.954
specificity 95%	Logistic	1,000,000	0.0092	stratified	-9.5e-06 (9.7e-06)	3.3e-09 (1.5e-10)	3.2e-09	7.8e-05	0.954	0.954

Table C.31: Bias, MSE, and Variability for specificity at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 99%	Logistic	5,000	0.0046	naive	-0.025 (0.013)	0.063 (0.0025)	0.062	0.39	0	0.988
specificity 99%	Logistic	5,000	0.0046	stratified	-0.012 (0.013)	0.066 (0.0023)	0.066	0.4	0	0.989
specificity 99%	Logistic	5,000	0.0092	naive	-0.026 (0.0037)	0.028 (0.0035)	0.027	0.098	0	0.99
specificity 99%	Logistic	5,000	0.0092	stratified	-0.023 (0.0036)	0.026 (0.0036)	0.025	0.097	0	0.991
specificity 99%	Logistic	5,000	0.0184	naive	-0.0047 (0.002)	0.0041 (0.0012)	0.0041	0.0018	0.198	0.994
specificity 99%	Logistic	5,000	0.0184	stratified	-0.0037 (0.0023)	0.0052 (0.0011)	0.0051	0.0018	0.318	0.994
specificity 99%	Logistic	10,000	0.0046	naive	-0.0044 (0.0013)	0.0015 (0.00052)	0.0015	0.0011	0.434	0.991
specificity 99%	Logistic	10,000	0.0046	stratified	-0.0029 (0.0013)	0.0016 (0.00058)	0.0016	0.0012	0.375	0.992
specificity 99%	Logistic	10,000	0.0092	naive	-0.00059 (0.00066)	0.00024 (0.00016)	0.00024	0.00091	0.595	0.992
specificity 99%	Logistic	10,000	0.0092	stratified	-0.0013 (0.00056)	0.00032 (0.00027)	0.00032	0.0011	0.467	0.993
specificity 99%	Logistic	10,000	0.0184	naive	-0.00022 (3.6e-05)	9.6e-07 (3.6e-07)	9.1e-07	0.001	0.973	0.995
specificity 99%	Logistic	10,000	0.0184	stratified	-0.00038 (0.00015)	2.3e-05 (2.3e-05)	2.3e-05	0.001	0.841	0.995
specificity 99%	Logistic	50,000	0.0046	naive	-2.3e-05 (1.6e-05)	6.3e-08 (2.9e-09)	6.2e-08	0.00034	0.99	0.992
specificity 99%	Logistic	50,000	0.0046	stratified	-2.7e-05 (1.6e-05)	6.2e-08 (2.7e-09)	6.1e-08	0.00034	0.99	0.992
specificity 99%	Logistic	50,000	0.0092	naive	-1.3e-05 (1.4e-05)	5.9e-08 (2.6e-09)	5.9e-08	0.00032	0.991	0.993
specificity 99%	Logistic	50,000	0.0092	stratified	-5.5e-05 (1.4e-05)	6.4e-08 (2.7e-09)	6.1e-08	0.00034	0.991	0.993
specificity 99%	Logistic	50,000	0.0184	naive	-1.6e-05 (1.2e-05)	6.4e-08 (2.9e-09)	6.4e-08	0.00035	0.993	0.994
specificity 99%	Logistic	50,000	0.0184	stratified	-1.9e-05 (1.3e-05)	7e-08 (3.1e-09)	6.9e-08	0.00035	0.993	0.994
specificity 99%	Logistic	100,000	0.0046	naive	-8.5e-06 (1.1e-05)	2.3e-08 (1e-09)	2.3e-08	0.0002	0.991	0.991
specificity 99%	Logistic	100,000	0.0046	stratified	-2.5e-05 (1.1e-05)	2.4e-08 (1.1e-09)	2.3e-08	0.0002	0.991	0.991
specificity 99%	Logistic	100,000	0.0092	naive	-2.5e-05 (9.7e-06)	2.7e-08 (1.3e-09)	2.7e-08	0.0002	0.991	0.993
specificity 99%	Logistic	100,000	0.0092	stratified	-1.4e-05 (1e-05)	2.6e-08 (1.1e-09)	2.6e-08	0.00022	0.992	0.993
specificity 99%	Logistic	100,000	0.0184	naive	-9.8e-06 (8.7e-06)	2.8e-08 (1.3e-09)	2.8e-08	0.00022	0.993	0.994
specificity 99%	Logistic	100,000	0.0184	stratified	-9.1e-06 (8.3e-06)	2.6e-08 (1.1e-09)	2.6e-08	0.00021	0.993	0.994
specificity 99%	Logistic	1,000,000	0.0046	naive	6.2e-07 (4.3e-06)	1.3e-09 (5.4e-11)	1.3e-09	4.7e-05	0.991	0.991
specificity 99%	Logistic	1,000,000	0.0046	stratified	-1.8e-08 (4.5e-06)	1.2e-09 (5.9e-11)	1.2e-09	4.6e-05	0.991	0.991
specificity 99%	Logistic	1,000,000	0.0092	naive	-3.1e-06 (4e-06)	1.4e-09 (6.2e-11)	1.4e-09	5.3e-05	0.992	0.992
specificity 99%	Logistic	1,000,000	0.0092	stratified	2.7e-06 (3.9e-06)	1.4e-09 (6.1e-11)	1.4e-09	5e-05	0.992	0.992

Table C.32: Bias, MSE, and Variability for specificity at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
specificity 99%	Logistic	5,000	0.0046	naive	-0.025 (0.013)	0.063 (0.0025)	0.062	0.39	0	0.988
specificity 99%	Logistic	5,000	0.0046	stratified	-0.012 (0.013)	0.066 (0.0023)	0.066	0.4	0	0.989
specificity 99%	Logistic	5,000	0.0092	naive	-0.026 (0.0037)	0.028 (0.0035)	0.027	0.098	0	0.99
specificity 99%	Logistic	5,000	0.0092	stratified	-0.023 (0.0036)	0.026 (0.0036)	0.025	0.097	0	0.991
specificity 99%	Logistic	5,000	0.0184	naive	-0.0047 (0.002)	0.0041 (0.0012)	0.0041	0.0018	0.198	0.994
specificity 99%	Logistic	5,000	0.0184	stratified	-0.0037 (0.0023)	0.0052 (0.0011)	0.0051	0.0018	0.318	0.994
specificity 99%	Logistic	10,000	0.0046	naive	-0.0044 (0.0013)	0.0015 (0.00052)	0.0015	0.0011	0.434	0.991
specificity 99%	Logistic	10,000	0.0046	stratified	-0.0029 (0.0013)	0.0016 (0.00058)	0.0016	0.0012	0.375	0.992
specificity 99%	Logistic	10,000	0.0092	naive	-0.00059 (0.00066)	0.00024 (0.00016)	0.00024	0.00091	0.595	0.992
specificity 99%	Logistic	10,000	0.0092	stratified	-0.0013 (0.00056)	0.00032 (0.00027)	0.00032	0.0011	0.467	0.993
specificity 99%	Logistic	10,000	0.0184	naive	-0.00022 (3.6e-05)	9.6e-07 (3.6e-07)	9.1e-07	0.001	0.973	0.995
specificity 99%	Logistic	10,000	0.0184	stratified	-0.00038 (0.00015)	2.3e-05 (2.3e-05)	2.3e-05	0.001	0.841	0.995
specificity 99%	Logistic	50,000	0.0046	naive	-2.3e-05 (1.6e-05)	6.3e-08 (2.9e-09)	6.2e-08	0.00034	0.99	0.992
specificity 99%	Logistic	50,000	0.0046	stratified	-2.7e-05 (1.6e-05)	6.2e-08 (2.7e-09)	6.1e-08	0.00034	0.99	0.992
specificity 99%	Logistic	50,000	0.0092	naive	-1.3e-05 (1.4e-05)	5.9e-08 (2.6e-09)	5.9e-08	0.00032	0.991	0.993
specificity 99%	Logistic	50,000	0.0092	stratified	-5.5e-05 (1.4e-05)	6.4e-08 (2.7e-09)	6.1e-08	0.00034	0.991	0.993
specificity 99%	Logistic	50,000	0.0184	naive	-1.6e-05 (1.2e-05)	6.4e-08 (2.9e-09)	6.4e-08	0.00035	0.993	0.994
specificity 99%	Logistic	50,000	0.0184	stratified	-1.9e-05 (1.3e-05)	7e-08 (3.1e-09)	6.9e-08	0.00035	0.993	0.994
specificity 99%	Logistic	100,000	0.0046	naive	-8.5e-06 (1.1e-05)	2.3e-08 (1e-09)	2.3e-08	0.0002	0.991	0.991
specificity 99%	Logistic	100,000	0.0046	stratified	-2.5e-05 (1.1e-05)	2.4e-08 (1.1e-09)	2.3e-08	0.0002	0.991	0.991
specificity 99%	Logistic	100,000	0.0092	naive	-2.5e-05 (9.7e-06)	2.7e-08 (1.3e-09)	2.7e-08	0.0002	0.991	0.993
specificity 99%	Logistic	100,000	0.0092	stratified	-1.4e-05 (1e-05)	2.6e-08 (1.1e-09)	2.6e-08	0.00022	0.992	0.993
specificity 99%	Logistic	100,000	0.0184	naive	-9.8e-06 (8.7e-06)	2.8e-08 (1.3e-09)	2.8e-08	0.00022	0.993	0.994
specificity 99%	Logistic	100,000	0.0184	stratified	-9.1e-06 (8.3e-06)	2.6e-08 (1.1e-09)	2.6e-08	0.00021	0.993	0.994
specificity 99%	Logistic	1,000,000	0.0046	naive	6.2e-07 (4.3e-06)	1.3e-09 (5.4e-11)	1.3e-09	4.7e-05	0.991	0.991
specificity 99%	Logistic	1,000,000	0.0046	stratified	-1.8e-08 (4.5e-06)	1.2e-09 (5.9e-11)	1.2e-09	4.6e-05	0.991	0.991
specificity 99%	Logistic	1,000,000	0.0092	naive	-3.1e-06 (4e-06)	1.4e-09 (6.2e-11)	1.4e-09	5.3e-05	0.992	0.992
specificity 99%	Logistic	1,000,000	0.0092	stratified	2.7e-06 (3.9e-06)	1.4e-09 (6.1e-11)	1.4e-09	5e-05	0.992	0.992

Table C.33: Bias, MSE, and Variability for specificity at the 99th percentile, logistic regression.

C.6 Negative Predictive Value (NPV)

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 90%	Random Forest	5,000	0.0046	naive	-0.00029 (3.2e-05)	6.2e-07 (4.5e-08)	5.3e-07	0.00097	0.994	1
NPV 90%	Random Forest	5,000	0.0046	stratified	-3.7e-05 (2.5e-05)	3.4e-07 (2.1e-08)	3.4e-07	0.00087	0.996	0.999
NPV 90%	Random Forest	5,000	0.0092	naive	-5.1e-05 (3.6e-05)	6.5e-07 (4e-08)	6.5e-07	0.0011	0.993	0.998
NPV 90%	Random Forest	5,000	0.0092	stratified	-6.4e-05 (3.4e-05)	6e-07 (3.8e-08)	6e-07	0.0011	0.994	0.998
NPV 90%	Random Forest	5,000	0.0184	naive	-8.8e-05 (4.7e-05)	1.1e-06 (6.2e-08)	1e-06	0.0014	0.989	0.994
NPV 90%	Random Forest	5,000	0.0184	stratified	-3.4e-05 (4.9e-05)	1.1e-06 (6.6e-08)	1.1e-06	0.0014	0.988	0.995
NPV 90%	Random Forest	10,000	0.0046	naive	-1.1e-05 (1.8e-05)	1.5e-07 (9e-09)	1.5e-07	0.00048	0.997	0.999
NPV 90%	Random Forest	10,000	0.0046	stratified	2.9e-06 (1.7e-05)	1.4e-07 (8.5e-09)	1.4e-07	0.00048	0.997	0.999
NPV 90%	Random Forest	10,000	0.0092	naive	-3.1e-05 (2.5e-05)	3.1e-07 (2e-08)	3.1e-07	0.00069	0.994	0.997
NPV 90%	Random Forest	10,000	0.0092	stratified	4.6e-06 (2.4e-05)	2.8e-07 (1.7e-08)	2.8e-07	0.00067	0.994	0.997
NPV 90%	Random Forest	10,000	0.0184	naive	-0.0001 (3.5e-05)	6e-07 (3.5e-08)	5.9e-07	0.0011	0.99	0.994
NPV 90%	Random Forest	10,000	0.0184	stratified	2.8e-05 (3.3e-05)	5.4e-07 (3.3e-08)	5.4e-07	0.00099	0.99	0.994
NPV 90%	Random Forest	50,000	0.0046	naive	-1.2e-05 (7.6e-06)	2.6e-08 (1.6e-09)	2.6e-08	0.00022	0.998	0.998
NPV 90%	Random Forest	50,000	0.0046	stratified	-1.4e-05 (7.4e-06)	2.5e-08 (1.5e-09)	2.5e-08	0.0002	0.997	0.998
NPV 90%	Random Forest	50,000	0.0092	naive	-4.1e-05 (1.1e-05)	5e-08 (3.2e-09)	4.8e-08	0.00031	0.995	0.997
NPV 90%	Random Forest	50,000	0.0092	stratified	-3.5e-05 (1.1e-05)	5.3e-08 (3.4e-09)	5.2e-08	0.00031	0.995	0.997
NPV 90%	Random Forest	50,000	0.0184	naive	-7.8e-05 (1.5e-05)	1e-07 (6.8e-09)	9.9e-08	0.00042	0.991	0.993
NPV 90%	Random Forest	50,000	0.0184	stratified	-4.4e-05 (1.6e-05)	1e-07 (6.2e-09)	1e-07	0.0004	0.991	0.993
NPV 90%	Random Forest	100,000	0.0046	naive	-1.6e-05 (5.5e-06)	1.3e-08 (8.1e-10)	1.3e-08	0.00015	0.998	0.998
NPV 90%	Random Forest	100,000	0.0046	stratified	-1.2e-05 (5.5e-06)	1.4e-08 (8.2e-10)	1.4e-08	0.00015	0.998	0.998
NPV 90%	Random Forest	100,000	0.0092	naive	-1.9e-05 (7.4e-06)	2.3e-08 (1.3e-09)	2.2e-08	0.0002	0.996	0.997
NPV 90%	Random Forest	100,000	0.0092	stratified	-1.1e-05 (8e-06)	2.6e-08 (1.7e-09)	2.6e-08	0.00021	0.996	0.997
NPV 90%	Random Forest	100,000	0.0184	naive	-3.1e-05 (1.1e-05)	4.9e-08 (3e-09)	4.8e-08	0.00031	0.991	0.993
NPV 90%	Random Forest	100,000	0.0184	stratified	-4.6e-05 (1.1e-05)	5e-08 (3e-09)	4.8e-08	0.0003	0.992	0.993
NPV 90%	Random Forest	1,000,000	0.0046	naive	-6.8e-06 (2.3e-06)	1.1e-09 (6.7e-11)	1e-09	4.4e-05	0.998	0.998
NPV 90%	Random Forest	1,000,000	0.0046	stratified	-6.5e-06 (2.3e-06)	1.1e-09 (6.8e-11)	1e-09	4.2e-05	0.998	0.998
NPV 90%	Random Forest	1,000,000	0.0092	naive	-1.4e-05 (3.1e-06)	1.8e-09 (1.1e-10)	1.6e-09	5.4e-05	0.996	0.996
NPV 90%	Random Forest	1,000,000	0.0092	stratified	-1.8e-05 (3.4e-06)	1.9e-09 (1.1e-10)	1.6e-09	5.3e-05	0.996	0.996

Table C.34: Bias, MSE, and Variability intervals for NPV at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 95%	Random Forest	5,000	0.0046	naive	-0.00037 (3.1e-05)	6.6e-07 (5e-08)	5.2e-07	0.00093	0.994	0.999
NPV 95%	Random Forest	5,000	0.0046	stratified	-4.3e-05 (2.4e-05)	3.1e-07 (1.9e-08)	3.1e-07	0.00084	0.995	0.999
NPV 95%	Random Forest	5,000	0.0092	naive	-7.5e-05 (3.4e-05)	6.1e-07 (4.4e-08)	6.1e-07	0.001	0.991	0.997
NPV 95%	Random Forest	5,000	0.0092	stratified	-3e-05 (3.3e-05)	5.5e-07 (3.8e-08)	5.5e-07	0.001	0.992	0.997
NPV 95%	Random Forest	5,000	0.0184	naive	-7.5e-05 (4.4e-05)	9.2e-07 (5.6e-08)	9.1e-07	0.0013	0.987	0.993
NPV 95%	Random Forest	5,000	0.0184	stratified	-5.9e-05 (4.5e-05)	9.8e-07 (5.9e-08)	9.8e-07	0.0013	0.986	0.992
NPV 95%	Random Forest	10,000	0.0046	naive	-1.3e-05 (1.7e-05)	1.3e-07 (8.7e-09)	1.3e-07	0.00043	0.996	0.998
NPV 95%	Random Forest	10,000	0.0046	stratified	-1.8e-05 (1.7e-05)	1.2e-07 (7.3e-09)	1.2e-07	0.00052	0.996	0.998
NPV 95%	Random Forest	10,000	0.0092	naive	-5.2e-05 (2.4e-05)	2.8e-07 (1.8e-08)	2.8e-07	0.00065	0.993	0.997
NPV 95%	Random Forest	10,000	0.0092	stratified	1.9e-06 (2.3e-05)	2.4e-07 (1.7e-08)	2.4e-07	0.00063	0.993	0.996
NPV 95%	Random Forest	10,000	0.0184	naive	-0.0001 (3.1e-05)	4.5e-07 (2.9e-08)	4.4e-07	0.00084	0.987	0.991
NPV 95%	Random Forest	10,000	0.0184	stratified	-2.8e-05 (3.2e-05)	4.6e-07 (2.9e-08)	4.6e-07	0.00085	0.988	0.992
NPV 95%	Random Forest	50,000	0.0046	naive	-1.5e-05 (7.2e-06)	2.2e-08 (1.4e-09)	2.2e-08	0.00017	0.997	0.998
NPV 95%	Random Forest	50,000	0.0046	stratified	-8.4e-06 (7.2e-06)	2.3e-08 (1.5e-09)	2.3e-08	0.00019	0.997	0.998
NPV 95%	Random Forest	50,000	0.0092	naive	-3.2e-05 (1.1e-05)	4.8e-08 (2.9e-09)	4.7e-08	0.0003	0.994	0.996
NPV 95%	Random Forest	50,000	0.0092	stratified	-3.6e-05 (1.1e-05)	5.2e-08 (3.4e-09)	5.1e-08	0.00029	0.994	0.996
NPV 95%	Random Forest	50,000	0.0184	naive	-7.9e-05 (1.5e-05)	9.4e-08 (6e-09)	8.8e-08	0.00038	0.989	0.991
NPV 95%	Random Forest	50,000	0.0184	stratified	-6.7e-05 (1.6e-05)	1e-07 (6.2e-09)	9.6e-08	0.00038	0.989	0.991
NPV 95%	Random Forest	100,000	0.0046	naive	-1.7e-05 (5.5e-06)	1.3e-08 (8.6e-10)	1.2e-08	0.00013	0.997	0.998
NPV 95%	Random Forest	100,000	0.0046	stratified	-1.8e-05 (5.4e-06)	1.3e-08 (8e-10)	1.3e-08	0.00016	0.997	0.998
NPV 95%	Random Forest	100,000	0.0092	naive	-1.8e-05 (6.9e-06)	2e-08 (1.2e-09)	2e-08	0.0002	0.995	0.995
NPV 95%	Random Forest	100,000	0.0092	stratified	-1.4e-05 (7.4e-06)	2.2e-08 (1.4e-09)	2.2e-08	0.0002	0.994	0.995
NPV 95%	Random Forest	100,000	0.0184	naive	-3.5e-05 (1.1e-05)	4.6e-08 (3e-09)	4.5e-08	0.00027	0.989	0.991
NPV 95%	Random Forest	100,000	0.0184	stratified	-3.5e-05 (1.2e-05)	4.8e-08 (2.8e-09)	4.7e-08	0.0003	0.989	0.99
NPV 95%	Random Forest	1,000,000	0.0046	naive	-6.6e-06 (2.3e-06)	1.1e-09 (6.6e-11)	1e-09	4.3e-05	0.997	0.998
NPV 95%	Random Forest	1,000,000	0.0046	stratified	-8.9e-06 (2.1e-06)	9.4e-10 (5.4e-11)	8.7e-10	3.9e-05	0.997	0.998
NPV 95%	Random Forest	1,000,000	0.0092	naive	-9.6e-06 (3.2e-06)	1.7e-09 (1.1e-10)	1.6e-09	5.6e-05	0.995	0.995
NPV 95%	Random Forest	1,000,000	0.0092	stratified	-1.1e-05 (3.2e-06)	1.6e-09 (1e-10)	1.5e-09	5.3e-05	0.995	0.995

Table C.35: Bias, MSE, and Variability for NPV at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 99%	Random Forest	5,000	0.0046	naive	-0.00045 (2.9e-05)	6.5e-07 (4.6e-08)	4.4e-07	0.00082	0.993	0.998
NPV 99%	Random Forest	5,000	0.0046	stratified	-1.9e-05 (1.9e-05)	1.9e-07 (1e-08)	1.9e-07	0.00061	0.995	0.997
NPV 99%	Random Forest	5,000	0.0092	naive	-8.4e-05 (2.7e-05)	3.7e-07 (2.3e-08)	3.6e-07	0.00081	0.991	0.994
NPV 99%	Random Forest	5,000	0.0092	stratified	-1.3e-05 (2.7e-05)	3.7e-07 (2e-08)	3.7e-07	0.00081	0.991	0.994
NPV 99%	Random Forest	5,000	0.0184	naive	-8.5e-05 (3.5e-05)	5.5e-07 (3.5e-08)	5.5e-07	0.001	0.982	0.987
NPV 99%	Random Forest	5,000	0.0184	stratified	-4.9e-05 (3.3e-05)	4.9e-07 (3.3e-08)	4.9e-07	0.00081	0.982	0.987
NPV 99%	Random Forest	10,000	0.0046	naive	-4.2e-05 (1.4e-05)	1.1e-07 (7e-09)	1e-07	0.00041	0.995	0.997
NPV 99%	Random Forest	10,000	0.0046	stratified	-2.3e-05 (1.3e-05)	9.3e-08 (5.8e-09)	9.2e-08	0.00041	0.995	0.997
NPV 99%	Random Forest	10,000	0.0092	naive	-5.4e-05 (1.8e-05)	1.7e-07 (1e-08)	1.6e-07	0.00053	0.991	0.994
NPV 99%	Random Forest	10,000	0.0092	stratified	-3.3e-06 (1.9e-05)	1.7e-07 (9.7e-09)	1.7e-07	0.0006	0.991	0.994
NPV 99%	Random Forest	10,000	0.0184	naive	-8.8e-05 (2.4e-05)	2.6e-07 (1.8e-08)	2.6e-07	0.00069	0.983	0.987
NPV 99%	Random Forest	10,000	0.0184	stratified	-6.1e-05 (2.5e-05)	2.6e-07 (1.7e-08)	2.6e-07	0.0007	0.983	0.986
NPV 99%	Random Forest	50,000	0.0046	naive	-1.3e-05 (6.1e-06)	1.6e-08 (1.1e-09)	1.6e-08	0.00016	0.996	0.997
NPV 99%	Random Forest	50,000	0.0046	stratified	-1.2e-05 (6.1e-06)	1.6e-08 (9.9e-10)	1.6e-08	0.00016	0.996	0.997
NPV 99%	Random Forest	50,000	0.0092	naive	-2.1e-05 (8.1e-06)	2.5e-08 (1.6e-09)	2.5e-08	0.0002	0.992	0.993
NPV 99%	Random Forest	50,000	0.0092	stratified	-3e-05 (8.3e-06)	2.5e-08 (1.6e-09)	2.4e-08	0.00018	0.992	0.993
NPV 99%	Random Forest	50,000	0.0184	naive	-4e-05 (1.1e-05)	3.9e-08 (2.4e-09)	3.7e-08	0.00026	0.985	0.986
NPV 99%	Random Forest	50,000	0.0184	stratified	-3.4e-05 (1.1e-05)	3.9e-08 (2.7e-09)	3.8e-08	0.00026	0.985	0.986
NPV 99%	Random Forest	100,000	0.0046	naive	-1e-05 (4.4e-06)	7.7e-09 (4.6e-10)	7.7e-09	0.00012	0.996	0.997
NPV 99%	Random Forest	100,000	0.0046	stratified	-1.5e-05 (4.4e-06)	7.8e-09 (5.2e-10)	7.6e-09	0.00011	0.996	0.997
NPV 99%	Random Forest	100,000	0.0092	naive	-1.8e-05 (6e-06)	1.3e-08 (7.7e-10)	1.2e-08	0.00015	0.992	0.993
NPV 99%	Random Forest	100,000	0.0092	stratified	1.2e-07 (5.9e-06)	1.2e-08 (7.3e-10)	1.2e-08	0.00014	0.993	0.993
NPV 99%	Random Forest	100,000	0.0184	naive	-1.8e-05 (7.8e-06)	1.8e-08 (1.1e-09)	1.8e-08	0.00018	0.985	0.986
NPV 99%	Random Forest	100,000	0.0184	stratified	-3.1e-05 (8.5e-06)	2e-08 (1.3e-09)	1.9e-08	0.00019	0.985	0.986
NPV 99%	Random Forest	1,000,000	0.0046	naive	-4.3e-06 (1.9e-06)	6.3e-10 (4.2e-11)	6.1e-10	3.1e-05	0.996	0.997
NPV 99%	Random Forest	1,000,000	0.0046	stratified	-2.4e-06 (1.8e-06)	5.2e-10 (3.1e-11)	5.2e-10	3e-05	0.996	0.997
NPV 99%	Random Forest	1,000,000	0.0092	naive	-5.4e-06 (2.6e-06)	8.8e-10 (5.2e-11)	8.5e-10	4.1e-05	0.993	0.993
NPV 99%	Random Forest	1,000,000	0.0092	stratified	-4.5e-06 (2.6e-06)	8.5e-10 (5.5e-11)	8.3e-10	3.8e-05	0.993	0.993

Table C.36: Bias, MSE, and Variability for NPV at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 90%	Ridge	5,000	0.0046	naive	-0.00024 (2.1e-05)	4.9e-07 (2.4e-08)	4.4e-07	0.00096	0.995	0.999
NPV 90%	Ridge	5,000	0.0046	stratified	-1.6e-05 (1.8e-05)	3.2e-07 (1.5e-08)	3.2e-07	0.00088	0.995	1
NPV 90%	Ridge	5,000	0.0092	naive	-2.7e-05 (2.5e-05)	6e-07 (2.7e-08)	6e-07	0.00096	0.993	0.998
NPV 90%	Ridge	5,000	0.0092	stratified	-2.9e-05 (2.4e-05)	5.8e-07 (2.5e-08)	5.8e-07	0.00092	0.993	0.998
NPV 90%	Ridge	5,000	0.0184	naive	-2.5e-05 (3.4e-05)	1.1e-06 (4.9e-08)	1.1e-06	0.0014	0.988	0.994
NPV 90%	Ridge	5,000	0.0184	stratified	-6.2e-05 (3.3e-05)	1.1e-06 (4.6e-08)	1.1e-06	0.0015	0.988	0.995
NPV 90%	Ridge	10,000	0.0046	naive	-2e-05 (1.3e-05)	1.6e-07 (6.8e-09)	1.6e-07	0.00056	0.997	0.999
NPV 90%	Ridge	10,000	0.0046	stratified	-2e-05 (1.2e-05)	1.4e-07 (6.4e-09)	1.4e-07	0.00054	0.997	0.999
NPV 90%	Ridge	10,000	0.0092	naive	-4.5e-05 (1.7e-05)	2.7e-07 (1.2e-08)	2.7e-07	0.00068	0.994	0.997
NPV 90%	Ridge	10,000	0.0092	stratified	-3.3e-05 (1.8e-05)	2.9e-07 (1.3e-08)	2.9e-07	0.00068	0.994	0.997
NPV 90%	Ridge	10,000	0.0184	naive	-8.6e-05 (2.4e-05)	5.6e-07 (2.4e-08)	5.5e-07	0.0011	0.989	0.994
NPV 90%	Ridge	10,000	0.0184	stratified	-5.7e-05 (2.5e-05)	5.7e-07 (2.6e-08)	5.7e-07	0.0011	0.989	0.994
NPV 90%	Ridge	50,000	0.0046	naive	-1.7e-05 (5.6e-06)	3e-08 (1.4e-09)	3e-08	0.00024	0.997	0.999
NPV 90%	Ridge	50,000	0.0046	stratified	-2.1e-05 (5.3e-06)	2.7e-08 (1.2e-09)	2.6e-08	0.00022	0.997	0.999
NPV 90%	Ridge	50,000	0.0092	naive	-2.3e-05 (7.4e-06)	5.1e-08 (2.2e-09)	5.1e-08	0.00029	0.995	0.997
NPV 90%	Ridge	50,000	0.0092	stratified	-2.7e-05 (7.6e-06)	5.3e-08 (2.4e-09)	5.2e-08	0.00033	0.995	0.997
NPV 90%	Ridge	50,000	0.0184	naive	-3.5e-05 (1.1e-05)	1e-07 (4.9e-09)	1e-07	0.0004	0.991	0.993
NPV 90%	Ridge	50,000	0.0184	stratified	-2.3e-05 (1.1e-05)	1.1e-07 (4.6e-09)	1.1e-07	0.00046	0.991	0.993
NPV 90%	Ridge	100,000	0.0046	naive	-1.3e-05 (3.9e-06)	1.4e-08 (6.2e-10)	1.4e-08	0.00015	0.998	0.998
NPV 90%	Ridge	100,000	0.0046	stratified	-6.9e-06 (4e-06)	1.5e-08 (6.8e-10)	1.5e-08	0.00016	0.998	0.998
NPV 90%	Ridge	100,000	0.0092	naive	-1.8e-05 (5.4e-06)	2.6e-08 (1.1e-09)	2.5e-08	0.00022	0.996	0.997
NPV 90%	Ridge	100,000	0.0092	stratified	-8.9e-06 (5.5e-06)	2.6e-08 (1.2e-09)	2.6e-08	0.00021	0.996	0.997
NPV 90%	Ridge	100,000	0.0184	naive	-2.3e-05 (7.6e-06)	4.7e-08 (2e-09)	4.7e-08	0.0003	0.992	0.993
NPV 90%	Ridge	100,000	0.0184	stratified	-1.3e-05 (7.7e-06)	4.8e-08 (2.3e-09)	4.8e-08	0.00028	0.992	0.993
NPV 90%	Ridge	1,000,000	0.0046	naive	1.5e-07 (1.6e-06)	1.1e-09 (4.9e-11)	1.1e-09	4.4e-05	0.998	0.998
NPV 90%	Ridge	1,000,000	0.0046	stratified	-2.5e-06 (1.6e-06)	1.2e-09 (4.7e-11)	1.1e-09	5e-05	0.998	0.998
NPV 90%	Ridge	1,000,000	0.0092	naive	-2e-06 (2.2e-06)	1.7e-09 (7.6e-11)	1.7e-09	5.2e-05	0.996	0.996
NPV 90%	Ridge	1,000,000	0.0092	stratified	-3.7e-06 (2.2e-06)	1.6e-09 (7.2e-11)	1.6e-09	5.7e-05	0.996	0.996

Table C.37: Bias, MSE, and Variability for NPV at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 95%	Ridge	5,000	0.0046	naive	-0.00029 (2e-05)	5.1e-07 (2.4e-08)	4.3e-07	0.00092	0.995	0.999
NPV 95%	Ridge	5,000	0.0046	stratified	-3.1e-05 (1.7e-05)	2.9e-07 (1.3e-08)	2.9e-07	0.00083	0.995	0.999
NPV 95%	Ridge	5,000	0.0092	naive	-5.7e-05 (2.4e-05)	5.6e-07 (2.5e-08)	5.6e-07	0.00089	0.992	0.997
NPV 95%	Ridge	5,000	0.0092	stratified	-2.3e-05 (2.3e-05)	5.3e-07 (2.4e-08)	5.3e-07	0.00087	0.992	0.997
NPV 95%	Ridge	5,000	0.0184	naive	-3.7e-05 (3.3e-05)	1e-06 (4.7e-08)	1e-06	0.0013	0.986	0.993
NPV 95%	Ridge	5,000	0.0184	stratified	-5.6e-05 (3.2e-05)	9.7e-07 (4.1e-08)	9.7e-07	0.0013	0.986	0.992
NPV 95%	Ridge	10,000	0.0046	naive	-2.6e-05 (1.2e-05)	1.5e-07 (6.7e-09)	1.5e-07	0.00052	0.996	0.998
NPV 95%	Ridge	10,000	0.0046	stratified	-2e-05 (1.1e-05)	1.3e-07 (5.8e-09)	1.3e-07	0.00052	0.996	0.999
NPV 95%	Ridge	10,000	0.0092	naive	-3.8e-05 (1.6e-05)	2.5e-07 (1.1e-08)	2.5e-07	0.00064	0.993	0.996
NPV 95%	Ridge	10,000	0.0092	stratified	-4.4e-05 (1.6e-05)	2.6e-07 (1.1e-08)	2.6e-07	0.00064	0.993	0.996
NPV 95%	Ridge	10,000	0.0184	naive	-7.7e-05 (2.4e-05)	5.3e-07 (2.4e-08)	5.3e-07	0.00099	0.987	0.992
NPV 95%	Ridge	10,000	0.0184	stratified	-4.1e-05 (2.3e-05)	4.7e-07 (2.1e-08)	4.7e-07	0.00094	0.987	0.992
NPV 95%	Ridge	50,000	0.0046	naive	-1.7e-05 (5.3e-06)	2.7e-08 (1.2e-09)	2.7e-08	0.00023	0.997	0.998
NPV 95%	Ridge	50,000	0.0046	stratified	-2.2e-05 (5.2e-06)	2.5e-08 (1.2e-09)	2.5e-08	0.00021	0.997	0.998
NPV 95%	Ridge	50,000	0.0092	naive	-2.8e-05 (7.2e-06)	4.9e-08 (2.2e-09)	4.8e-08	0.00028	0.994	0.996
NPV 95%	Ridge	50,000	0.0092	stratified	-2.1e-05 (7.3e-06)	4.7e-08 (2.1e-09)	4.7e-08	0.00031	0.994	0.996
NPV 95%	Ridge	50,000	0.0184	naive	-3.5e-05 (1e-05)	9.9e-08 (4.7e-09)	9.8e-08	0.0004	0.989	0.991
NPV 95%	Ridge	50,000	0.0184	stratified	-2.8e-05 (1e-05)	9.2e-08 (4.1e-09)	9.1e-08	0.0004	0.989	0.991
NPV 95%	Ridge	100,000	0.0046	naive	-1.2e-05 (3.8e-06)	1.3e-08 (5.8e-10)	1.3e-08	0.00016	0.997	0.998
NPV 95%	Ridge	100,000	0.0046	stratified	-9.4e-06 (3.9e-06)	1.4e-08 (6.1e-10)	1.4e-08	0.00016	0.997	0.998
NPV 95%	Ridge	100,000	0.0092	naive	-2e-05 (5.3e-06)	2.5e-08 (1.1e-09)	2.4e-08	0.00021	0.995	0.996
NPV 95%	Ridge	100,000	0.0092	stratified	-1.5e-05 (5.2e-06)	2.4e-08 (1.1e-09)	2.4e-08	0.0002	0.995	0.996
NPV 95%	Ridge	100,000	0.0184	naive	-2.1e-05 (7.2e-06)	4.2e-08 (1.8e-09)	4.2e-08	0.00028	0.989	0.991
NPV 95%	Ridge	100,000	0.0184	stratified	-1.7e-05 (7.6e-06)	4.7e-08 (2.3e-09)	4.7e-08	0.00028	0.989	0.991
NPV 95%	Ridge	1,000,000	0.0046	naive	2.5e-07 (1.6e-06)	1.1e-09 (4.7e-11)	1.1e-09	4.3e-05	0.998	0.998
NPV 95%	Ridge	1,000,000	0.0046	stratified	-1.1e-06 (1.6e-06)	1.1e-09 (4.9e-11)	1.1e-09	4.5e-05	0.998	0.998
NPV 95%	Ridge	1,000,000	0.0092	naive	-2.9e-06 (2.2e-06)	1.6e-09 (7.1e-11)	1.6e-09	5.4e-05	0.995	0.995
NPV 95%	Ridge	1,000,000	0.0092	stratified	-1.3e-06 (2.1e-06)	1.6e-09 (7.3e-11)	1.6e-09	5.1e-05	0.995	0.995

Table C.38: Bias, MSE, and Variability for NPV at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 99%	Ridge	5,000	0.0046	naive	-0.00038 (1.9e-05)	5.3e-07 (2.7e-08)	3.8e-07	0.00085	0.993	0.997
NPV 99%	Ridge	5,000	0.0046	stratified	-3.5e-05 (1.3e-05)	1.8e-07 (7.4e-09)	1.8e-07	0.00061	0.995	0.998
NPV 99%	Ridge	5,000	0.0092	naive	-9.4e-05 (2e-05)	3.9e-07 (1.7e-08)	3.8e-07	0.00081	0.991	0.994
NPV 99%	Ridge	5,000	0.0092	stratified	-3e-05 (1.9e-05)	3.3e-07 (1.4e-08)	3.3e-07	0.00067	0.991	0.995
NPV 99%	Ridge	5,000	0.0184	naive	2.1e-05 (2.5e-05)	5e-07 (2.3e-08)	5e-07	0.00083	0.983	0.987
NPV 99%	Ridge	5,000	0.0184	stratified	-6.2e-05 (2.4e-05)	5.1e-07 (2.2e-08)	5.1e-07	0.001	0.983	0.987
NPV 99%	Ridge	10,000	0.0046	naive	-3.2e-05 (9.7e-06)	9.5e-08 (4.3e-09)	9.4e-08	0.0004	0.995	0.997
NPV 99%	Ridge	10,000	0.0046	stratified	-1.6e-05 (9.5e-06)	9.2e-08 (4e-09)	9.2e-08	0.0004	0.996	0.997
NPV 99%	Ridge	10,000	0.0092	naive	-3.2e-05 (1.3e-05)	1.7e-07 (6.8e-09)	1.7e-07	0.00051	0.992	0.994
NPV 99%	Ridge	10,000	0.0092	stratified	-1.7e-05 (1.3e-05)	1.7e-07 (7.5e-09)	1.7e-07	0.00051	0.992	0.994
NPV 99%	Ridge	10,000	0.0184	naive	-2.5e-05 (1.8e-05)	2.7e-07 (1.2e-08)	2.7e-07	0.0007	0.983	0.987
NPV 99%	Ridge	10,000	0.0184	stratified	-5.9e-05 (1.8e-05)	2.6e-07 (1.1e-08)	2.6e-07	0.0007	0.984	0.987
NPV 99%	Ridge	50,000	0.0046	naive	-1.7e-05 (4.3e-06)	1.6e-08 (7.4e-10)	1.6e-08	0.00016	0.996	0.997
NPV 99%	Ridge	50,000	0.0046	stratified	-2.5e-05 (4.2e-06)	1.7e-08 (7.3e-10)	1.7e-08	0.00018	0.996	0.997
NPV 99%	Ridge	50,000	0.0092	naive	-2.1e-05 (5.8e-06)	2.9e-08 (1.3e-09)	2.8e-08	0.00022	0.992	0.993
NPV 99%	Ridge	50,000	0.0092	stratified	-2e-05 (6.1e-06)	3.2e-08 (1.4e-09)	3.1e-08	0.00024	0.992	0.993
NPV 99%	Ridge	50,000	0.0184	naive	-3.3e-05 (7.9e-06)	4.7e-08 (2.1e-09)	4.6e-08	0.00028	0.985	0.986
NPV 99%	Ridge	50,000	0.0184	stratified	-3.5e-05 (7.8e-06)	4.5e-08 (2.2e-09)	4.4e-08	0.00028	0.985	0.986
NPV 99%	Ridge	100,000	0.0046	naive	-1.1e-05 (3.1e-06)	8.2e-09 (3.5e-10)	8.1e-09	0.00012	0.996	0.997
NPV 99%	Ridge	100,000	0.0046	stratified	-9.3e-06 (3.1e-06)	8.4e-09 (3.5e-10)	8.3e-09	0.00012	0.996	0.997
NPV 99%	Ridge	100,000	0.0092	naive	-2.1e-05 (4.4e-06)	1.5e-08 (6.5e-10)	1.5e-08	0.00017	0.993	0.993
NPV 99%	Ridge	100,000	0.0092	stratified	-1.6e-05 (4.4e-06)	1.5e-08 (6.6e-10)	1.5e-08	0.00017	0.993	0.993
NPV 99%	Ridge	100,000	0.0184	naive	-1.3e-05 (5.8e-06)	2.2e-08 (1e-09)	2.2e-08	0.00019	0.985	0.986
NPV 99%	Ridge	100,000	0.0184	stratified	-1.9e-05 (5.8e-06)	2.3e-08 (1e-09)	2.2e-08	0.0002	0.985	0.986
NPV 99%	Ridge	1,000,000	0.0046	naive	-1e-06 (1.4e-06)	7.4e-10 (3.4e-11)	7.4e-10	3.6e-05	0.996	0.997
NPV 99%	Ridge	1,000,000	0.0046	stratified	2.3e-07 (1.3e-06)	7.2e-10 (3.1e-11)	7.2e-10	3.7e-05	0.996	0.997
NPV 99%	Ridge	1,000,000	0.0092	naive	-1.7e-06 (1.8e-06)	9.5e-10 (4.3e-11)	9.4e-10	3.9e-05	0.993	0.993
NPV 99%	Ridge	1,000,000	0.0092	stratified	-2e-06 (1.8e-06)	9.9e-10 (4.4e-11)	9.9e-10	4.2e-05	0.993	0.993

Table C.39: Bias, MSE, and Variability for NPV at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 90%	Logistic	5,000	0.0046	naive	0.00019 (3.5e-05)	3.3e-06 (4.1e-07)	3.2e-06	0.0017	0.976	1
NPV 90%	Logistic	5,000	0.0046	stratified	0.00019 (2.1e-05)	1.4e-06 (1.1e-07)	1.4e-06	0.00082	0.987	1
NPV 90%	Logistic	5,000	0.0092	naive	-0.00035 (3.2e-05)	1.8e-06 (4.1e-07)	1.6e-06	0.0011	0.973	0.998
NPV 90%	Logistic	5,000	0.0092	stratified	-0.00018 (2.9e-05)	9e-07 (1e-07)	8.5e-07	0.001	0.985	0.996
NPV 90%	Logistic	5,000	0.0184	naive	-0.00039 (4.2e-05)	2.8e-06 (1.1e-06)	2.6e-06	0.0016	0.955	0.991
NPV 90%	Logistic	5,000	0.0184	stratified	-0.00041 (4.3e-05)	1.9e-06 (1.3e-07)	1.7e-06	0.0015	0.98	0.993
NPV 90%	Logistic	10,000	0.0046	naive	-0.00017 (1.4e-05)	2e-07 (1.5e-08)	1.8e-07	0.00052	0.993	0.998
NPV 90%	Logistic	10,000	0.0046	stratified	-9.4e-05 (1.3e-05)	1.5e-07 (6.2e-09)	1.4e-07	0.00054	0.995	0.997
NPV 90%	Logistic	10,000	0.0092	naive	-0.00024 (2.1e-05)	4.2e-07 (1.8e-08)	3.7e-07	0.00079	0.992	0.996
NPV 90%	Logistic	10,000	0.0092	stratified	-0.00023 (2e-05)	3.8e-07 (1.7e-08)	3.3e-07	0.00077	0.992	0.996
NPV 90%	Logistic	10,000	0.0184	naive	-0.0003 (2.7e-05)	7.3e-07 (3.2e-08)	6.4e-07	0.0011	0.987	0.992
NPV 90%	Logistic	10,000	0.0184	stratified	-0.00033 (2.7e-05)	7.7e-07 (3.6e-08)	6.6e-07	0.0011	0.986	0.992
NPV 90%	Logistic	50,000	0.0046	naive	-5.9e-05 (5.8e-06)	3.4e-08 (1.5e-09)	3.1e-08	0.00022	0.997	0.998
NPV 90%	Logistic	50,000	0.0046	stratified	-5.7e-05 (6e-06)	3.5e-08 (1.6e-09)	3.2e-08	0.00022	0.997	0.998
NPV 90%	Logistic	50,000	0.0092	naive	-6.7e-05 (8.2e-06)	6.7e-08 (3e-09)	6.3e-08	0.00033	0.995	0.997
NPV 90%	Logistic	50,000	0.0092	stratified	-4.5e-05 (7.8e-06)	5.8e-08 (2.6e-09)	5.6e-08	0.00033	0.995	0.997
NPV 90%	Logistic	50,000	0.0184	naive	-6.8e-05 (1.1e-05)	1.2e-07 (5.5e-09)	1.1e-07	0.00045	0.991	0.993
NPV 90%	Logistic	50,000	0.0184	stratified	-5.9e-05 (1.1e-05)	1.1e-07 (5.1e-09)	1.1e-07	0.00045	0.991	0.993
NPV 90%	Logistic	100,000	0.0046	naive	-2.6e-05 (4.2e-06)	1.7e-08 (7.9e-10)	1.6e-08	0.00017	0.998	0.998
NPV 90%	Logistic	100,000	0.0046	stratified	-2.8e-05 (4.1e-06)	1.6e-08 (6.9e-10)	1.5e-08	0.00017	0.998	0.998
NPV 90%	Logistic	100,000	0.0092	naive	-2.6e-05 (5.5e-06)	2.8e-08 (1.3e-09)	2.7e-08	0.00022	0.996	0.997
NPV 90%	Logistic	100,000	0.0092	stratified	-3.4e-05 (5.7e-06)	3e-08 (1.4e-09)	2.9e-08	0.00024	0.995	0.997
NPV 90%	Logistic	100,000	0.0184	naive	-2.1e-05 (7.6e-06)	5.1e-08 (2.2e-09)	5.1e-08	0.0003	0.992	0.993
NPV 90%	Logistic	100,000	0.0184	stratified	-3.6e-05 (7.6e-06)	4.9e-08 (2.1e-09)	4.8e-08	0.00031	0.992	0.993
NPV 90%	Logistic	1,000,000	0.0046	naive	-1.7e-06 (1.6e-06)	1.1e-09 (4.9e-11)	1.1e-09	4.8e-05	0.998	0.998
NPV 90%	Logistic	1,000,000	0.0046	stratified	-1.1e-06 (1.6e-06)	1.1e-09 (4.6e-11)	1.1e-09	4.7e-05	0.998	0.998
NPV 90%	Logistic	1,000,000	0.0092	naive	-9.6e-07 (2.3e-06)	1.7e-09 (8.1e-11)	1.7e-09	5.7e-05	0.996	0.996
NPV 90%	Logistic	1,000,000	0.0092	stratified	1.7e-07 (2.3e-06)	1.7e-09 (7.8e-11)	1.7e-09	5.8e-05	0.996	0.996

Table C.40: Bias, MSE, and Variability for NPV at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 95%	Logistic	5,000	0.0046	naive	0.00023 (3.4e-05)	2.9e-06 (4.1e-07)	2.9e-06	0.0016	0.976	1
NPV 95%	Logistic	5,000	0.0046	stratified	0.00022 (1.8e-05)	1.2e-06 (8e-08)	1.1e-06	0.00076	0.99	1
NPV 95%	Logistic	5,000	0.0092	naive	-0.00026 (2.6e-05)	1.3e-06 (1.8e-07)	1.2e-06	0.001	0.983	0.998
NPV 95%	Logistic	5,000	0.0092	stratified	-0.00013 (2.5e-05)	6.2e-07 (7.6e-08)	5.9e-07	0.00084	0.986	0.995
NPV 95%	Logistic	5,000	0.0184	naive	-0.00028 (3.6e-05)	2.3e-06 (1e-06)	2.2e-06	0.0014	0.955	0.991
NPV 95%	Logistic	5,000	0.0184	stratified	-0.00031 (3.7e-05)	1.3e-06 (7.7e-08)	1.2e-06	0.0014	0.981	0.991
NPV 95%	Logistic	10,000	0.0046	naive	-0.00015 (1.2e-05)	1.6e-07 (1.5e-08)	1.3e-07	0.00051	0.993	0.997
NPV 95%	Logistic	10,000	0.0046	stratified	-6.8e-05 (1.1e-05)	1e-07 (4.1e-09)	9.7e-08	0.00042	0.995	0.997
NPV 95%	Logistic	10,000	0.0092	naive	-0.00019 (1.7e-05)	3.2e-07 (1.4e-08)	2.8e-07	0.00072	0.991	0.995
NPV 95%	Logistic	10,000	0.0092	stratified	-0.00019 (1.7e-05)	2.9e-07 (1.3e-08)	2.5e-07	0.00064	0.992	0.995
NPV 95%	Logistic	10,000	0.0184	naive	-0.00023 (2.4e-05)	5.8e-07 (2.4e-08)	5.3e-07	0.00097	0.986	0.99
NPV 95%	Logistic	10,000	0.0184	stratified	-0.00027 (2.3e-05)	5.8e-07 (2.6e-08)	5.1e-07	0.00096	0.986	0.99
NPV 95%	Logistic	50,000	0.0046	naive	-5.4e-05 (5.3e-06)	3.1e-08 (1.3e-09)	2.8e-08	0.00021	0.997	0.998
NPV 95%	Logistic	50,000	0.0046	stratified	-5.3e-05 (5.5e-06)	3.1e-08 (1.4e-09)	2.9e-08	0.00021	0.997	0.998
NPV 95%	Logistic	50,000	0.0092	naive	-6e-05 (7.7e-06)	5.8e-08 (2.8e-09)	5.5e-08	0.00031	0.994	0.996
NPV 95%	Logistic	50,000	0.0092	stratified	-4.5e-05 (7.4e-06)	5.3e-08 (2.4e-09)	5.1e-08	0.00032	0.994	0.996
NPV 95%	Logistic	50,000	0.0184	naive	-6.7e-05 (1e-05)	1e-07 (4.4e-09)	9.7e-08	0.00044	0.989	0.991
NPV 95%	Logistic	50,000	0.0184	stratified	-4.9e-05 (1.1e-05)	1e-07 (4.7e-09)	1e-07	0.00042	0.989	0.991
NPV 95%	Logistic	100,000	0.0046	naive	-2.5e-05 (4e-06)	1.6e-08 (6.8e-10)	1.5e-08	0.00016	0.997	0.998
NPV 95%	Logistic	100,000	0.0046	stratified	-2.6e-05 (4e-06)	1.5e-08 (6.5e-10)	1.4e-08	0.00016	0.997	0.998
NPV 95%	Logistic	100,000	0.0092	naive	-2.7e-05 (5.5e-06)	2.8e-08 (1.2e-09)	2.7e-08	0.00022	0.995	0.995
NPV 95%	Logistic	100,000	0.0092	stratified	-3.1e-05 (5.4e-06)	2.7e-08 (1.2e-09)	2.6e-08	0.0002	0.994	0.995
NPV 95%	Logistic	100,000	0.0184	naive	-1.7e-05 (7.4e-06)	4.7e-08 (2e-09)	4.7e-08	0.0003	0.989	0.991
NPV 95%	Logistic	100,000	0.0184	stratified	-2.5e-05 (7.3e-06)	4.3e-08 (1.8e-09)	4.3e-08	0.00029	0.989	0.99
NPV 95%	Logistic	1,000,000	0.0046	naive	-6.7e-07 (1.6e-06)	1.1e-09 (4.9e-11)	1.1e-09	4.5e-05	0.998	0.998
NPV 95%	Logistic	1,000,000	0.0046	stratified	-8.8e-07 (1.6e-06)	1.1e-09 (4.9e-11)	1.1e-09	4.4e-05	0.998	0.998
NPV 95%	Logistic	1,000,000	0.0092	naive	6.7e-07 (2.2e-06)	1.7e-09 (8.2e-11)	1.7e-09	5.5e-05	0.995	0.995
NPV 95%	Logistic	1,000,000	0.0092	stratified	1.2e-06 (2.3e-06)	1.7e-09 (7.7e-11)	1.7e-09	5.7e-05	0.995	0.995

Table C.41: Bias, MSE, and Variability for NPV at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
NPV 99%	Logistic	5,000	0.0046	naive	-0.00042 (2.9e-05)	1.8e-06 (1.5e-07)	1.4e-06	0.0012	0.987	0.999
NPV 99%	Logistic	5,000	0.0046	stratified	4.8e-05 (1.5e-05)	3.8e-07 (5.2e-08)	3.8e-07	0.00063	0.99	0.998
NPV 99%	Logistic	5,000	0.0092	naive	-0.00012 (2.2e-05)	8.6e-07 (1.2e-07)	8.4e-07	0.00067	0.983	0.998
NPV 99%	Logistic	5,000	0.0092	stratified	-3.9e-05 (1.8e-05)	3.4e-07 (5.1e-08)	3.4e-07	0.00061	0.986	0.994
NPV 99%	Logistic	5,000	0.0184	naive	-9.8e-05 (2.4e-05)	6.8e-07 (1.3e-07)	6.7e-07	0.00099	0.973	0.987
NPV 99%	Logistic	5,000	0.0184	stratified	-0.00011 (2.5e-05)	5.2e-07 (3e-08)	5.1e-07	0.00098	0.981	0.988
NPV 99%	Logistic	10,000	0.0046	naive	-9.9e-05 (8.3e-06)	8.5e-08 (9.4e-09)	7.5e-08	0.0003	0.993	0.997
NPV 99%	Logistic	10,000	0.0046	stratified	-3.2e-05 (7.5e-06)	4.8e-08 (2.1e-09)	4.7e-08	0.0003	0.995	0.997
NPV 99%	Logistic	10,000	0.0092	naive	-8.5e-05 (1.1e-05)	1.3e-07 (5.8e-09)	1.2e-07	0.0005	0.991	0.993
NPV 99%	Logistic	10,000	0.0092	stratified	-8.8e-05 (1.2e-05)	1.4e-07 (6.1e-09)	1.3e-07	0.0005	0.991	0.993
NPV 99%	Logistic	10,000	0.0184	naive	-0.00011 (1.7e-05)	2.6e-07 (1.2e-08)	2.5e-07	0.00069	0.983	0.986
NPV 99%	Logistic	10,000	0.0184	stratified	-0.00015 (1.7e-05)	2.6e-07 (1.1e-08)	2.4e-07	0.00061	0.983	0.986
NPV 99%	Logistic	50,000	0.0046	naive	-3.2e-05 (4.1e-06)	1.7e-08 (7.1e-10)	1.6e-08	0.00018	0.996	0.997
NPV 99%	Logistic	50,000	0.0046	stratified	-3.6e-05 (4.2e-06)	1.8e-08 (7.6e-10)	1.7e-08	0.00018	0.996	0.997
NPV 99%	Logistic	50,000	0.0092	naive	-4.8e-05 (6e-06)	3.4e-08 (1.5e-09)	3.2e-08	0.00024	0.992	0.993
NPV 99%	Logistic	50,000	0.0092	stratified	-3.6e-05 (6e-06)	3.1e-08 (1.4e-09)	3e-08	0.00022	0.992	0.993
NPV 99%	Logistic	50,000	0.0184	naive	-4.5e-05 (8.1e-06)	5e-08 (2.1e-09)	4.8e-08	0.0003	0.985	0.986
NPV 99%	Logistic	50,000	0.0184	stratified	-3.9e-05 (8.1e-06)	5.1e-08 (2.3e-09)	4.9e-08	0.00029	0.985	0.986
NPV 99%	Logistic	100,000	0.0046	naive	-2e-05 (3.1e-06)	9.1e-09 (4e-10)	8.7e-09	0.00012	0.996	0.997
NPV 99%	Logistic	100,000	0.0046	stratified	-1.8e-05 (3.1e-06)	9.3e-09 (4e-10)	9e-09	0.00013	0.996	0.997
NPV 99%	Logistic	100,000	0.0092	naive	-2.2e-05 (4.4e-06)	1.6e-08 (7.1e-10)	1.6e-08	0.00017	0.992	0.993
NPV 99%	Logistic	100,000	0.0092	stratified	-3e-05 (4.3e-06)	1.6e-08 (7.5e-10)	1.5e-08	0.00016	0.992	0.993
NPV 99%	Logistic	100,000	0.0184	naive	-7.8e-06 (5.7e-06)	2.3e-08 (1.1e-09)	2.3e-08	0.0002	0.985	0.986
NPV 99%	Logistic	100,000	0.0184	stratified	-2.3e-05 (5.7e-06)	2.3e-08 (1e-09)	2.2e-08	0.00019	0.985	0.986
NPV 99%	Logistic	1,000,000	0.0046	naive	-1.1e-06 (1.4e-06)	7.5e-10 (3.3e-11)	7.5e-10	3.6e-05	0.996	0.997
NPV 99%	Logistic	1,000,000	0.0046	stratified	-2.9e-06 (1.3e-06)	7.3e-10 (3.3e-11)	7.2e-10	3.6e-05	0.996	0.997
NPV 99%	Logistic	1,000,000	0.0092	naive	1.3e-06 (1.8e-06)	1.1e-09 (5.2e-11)	1.1e-09	4.6e-05	0.993	0.993
NPV 99%	Logistic	1,000,000	0.0092	stratified	1.1e-06 (1.8e-06)	1.1e-09 (4.8e-11)	1.1e-09	4.3e-05	0.993	0.993

Table C.42: Bias, MSE, and Variability for NPV at the 99th percentile, logistic regression.

C.7 Positive Predictive Value (PPV)

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 90%	Random Forest	5,000	0.0046	naive	0.0025 (0.00027)	4.6e-05 (3.6e-06)	4e-05	0.0081	0.0108	0.0571
PPV 90%	Random Forest	5,000	0.0046	stratified	-2.3e-05 (0.00022)	2.6e-05 (1.6e-06)	2.6e-05	0.0068	0.00896	0.0391
PPV 90%	Random Forest	5,000	0.0092	naive	0.00043 (0.00032)	5.2e-05 (3e-06)	5.2e-05	0.0098	0.0314	0.0684
PPV 90%	Random Forest	5,000	0.0092	stratified	5.9e-06 (0.0003)	4.8e-05 (3e-06)	4.8e-05	0.0091	0.0305	0.0715
PPV 90%	Random Forest	5,000	0.0184	naive	-0.00031 (0.00044)	8.3e-05 (5e-06)	8.3e-05	0.013	0.0797	0.131
PPV 90%	Random Forest	5,000	0.0184	stratified	0.001 (0.00043)	8.4e-05 (5.1e-06)	8.3e-05	0.013	0.0751	0.132
PPV 90%	Random Forest	10,000	0.0046	naive	0.00043 (0.00017)	1.3e-05 (7.6e-07)	1.3e-05	0.0051	0.0167	0.0356
PPV 90%	Random Forest	10,000	0.0046	stratified	0.00018 (0.00016)	1.1e-05 (7e-07)	1.1e-05	0.0045	0.0155	0.0358
PPV 90%	Random Forest	10,000	0.0092	naive	-9e-05 (0.00022)	2.4e-05 (1.5e-06)	2.4e-05	0.0064	0.0379	0.0657
PPV 90%	Random Forest	10,000	0.0092	stratified	0.00039 (0.00021)	2.1e-05 (1.3e-06)	2.1e-05	0.0061	0.0399	0.0662
PPV 90%	Random Forest	10,000	0.0184	naive	-0.00058 (0.00032)	4.6e-05 (2.8e-06)	4.6e-05	0.0094	0.0872	0.127
PPV 90%	Random Forest	10,000	0.0184	stratified	0.0011 (0.0003)	4.2e-05 (2.7e-06)	4e-05	0.0083	0.0898	0.126
PPV 90%	Random Forest	50,000	0.0046	naive	-1.8e-05 (6.8e-05)	2e-06 (1.3e-07)	2e-06	0.002	0.0233	0.0315
PPV 90%	Random Forest	50,000	0.0046	stratified	-2.2e-05 (6.7e-05)	2e-06 (1.2e-07)	2e-06	0.0019	0.0229	0.032
PPV 90%	Random Forest	50,000	0.0092	naive	-0.00014 (9.8e-05)	3.9e-06 (2.5e-07)	3.9e-06	0.0028	0.0498	0.0624
PPV 90%	Random Forest	50,000	0.0092	stratified	-4.7e-05 (0.0001)	4.1e-06 (2.6e-07)	4.1e-06	0.0027	0.0503	0.0625
PPV 90%	Random Forest	50,000	0.0184	naive	-0.00042 (0.00014)	8.1e-06 (5.1e-07)	7.9e-06	0.0037	0.101	0.119
PPV 90%	Random Forest	50,000	0.0184	stratified	-0.00016 (0.00014)	8.2e-06 (4.9e-07)	8.2e-06	0.0037	0.103	0.119
PPV 90%	Random Forest	100,000	0.0046	naive	-4.7e-05 (5e-05)	1e-06 (6.5e-08)	1e-06	0.0013	0.0252	0.0317
PPV 90%	Random Forest	100,000	0.0046	stratified	-1.8e-05 (5e-05)	1.1e-06 (6.7e-08)	1.1e-06	0.0013	0.0257	0.0317
PPV 90%	Random Forest	100,000	0.0092	naive	-3.3e-05 (6.8e-05)	1.8e-06 (1.1e-07)	1.8e-06	0.0018	0.0527	0.0604
PPV 90%	Random Forest	100,000	0.0092	stratified	1.4e-05 (7.2e-05)	2.1e-06 (1.3e-07)	2.1e-06	0.0019	0.053	0.0625
PPV 90%	Random Forest	100,000	0.0184	naive	-0.00021 (0.0001)	3.9e-06 (2.5e-07)	3.9e-06	0.0026	0.106	0.118
PPV 90%	Random Forest	100,000	0.0184	stratified	-0.00025 (0.0001)	3.8e-06 (2.4e-07)	3.8e-06	0.0027	0.107	0.119
PPV 90%	Random Forest	1,000,000	0.0046	naive	-6.1e-05 (2.1e-05)	8.9e-08 (5.5e-09)	8.5e-08	0.00039	0.0281	0.0298
PPV 90%	Random Forest	1,000,000	0.0046	stratified	-6.3e-05 (2.1e-05)	8.6e-08 (5.5e-09)	8.2e-08	0.00039	0.0279	0.0299
PPV 90%	Random Forest	1,000,000	0.0092	naive	-0.00013 (3e-05)	1.5e-07 (9.1e-09)	1.3e-07	0.0005	0.0563	0.0585
PPV 90%	Random Forest	1,000,000	0.0092	stratified	-0.00016 (3.1e-05)	1.5e-07 (8.6e-09)	1.2e-07	0.00046	0.0564	0.0583

Table C.43: Bias, MSE, and Variability for PPV at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 95%	Random Forest	5,000	0.0046	naive	0.0032 (0.00053)	0.00016 (1.3e-05)	0.00015	0.015	0.0111	0.0963
PPV 95%	Random Forest	5,000	0.0046	stratified	-0.00034 (0.00044)	0.00011 (6.5e-06)	0.00011	0.015	0.0037	0.0677
PPV 95%	Random Forest	5,000	0.0092	naive	0.00037 (0.00064)	0.0002 (1.4e-05)	0.0002	0.018	0.0322	0.121
PPV 95%	Random Forest	5,000	0.0092	stratified	0.00051 (0.00062)	0.00019 (1.3e-05)	0.00019	0.018	0.0282	0.123
PPV 95%	Random Forest	5,000	0.0184	naive	-0.00023 (0.00089)	0.00033 (2.1e-05)	0.00033	0.025	0.106	0.217
PPV 95%	Random Forest	5,000	0.0184	stratified	0.0019 (0.00087)	0.00035 (2.1e-05)	0.00035	0.026	0.108	0.209
PPV 95%	Random Forest	10,000	0.0046	naive	0.00098 (0.00033)	5e-05 (3e-06)	4.9e-05	0.009	0.0191	0.0596
PPV 95%	Random Forest	10,000	0.0046	stratified	7.4e-05 (0.00033)	4.7e-05 (2.8e-06)	4.7e-05	0.0097	0.0194	0.0587
PPV 95%	Random Forest	10,000	0.0092	naive	-0.00062 (0.00044)	9.7e-05 (6.4e-06)	9.7e-05	0.012	0.0527	0.116
PPV 95%	Random Forest	10,000	0.0092	stratified	0.00098 (0.00042)	8.5e-05 (5.6e-06)	8.4e-05	0.012	0.0539	0.111
PPV 95%	Random Forest	10,000	0.0184	naive	-0.00095 (0.00057)	0.00015 (9.4e-06)	0.00015	0.015	0.13	0.201
PPV 95%	Random Forest	10,000	0.0184	stratified	0.00089 (0.00058)	0.00015 (9.7e-06)	0.00015	0.017	0.135	0.208
PPV 95%	Random Forest	50,000	0.0046	naive	-5.7e-05 (0.00014)	7.8e-06 (4.8e-07)	7.8e-06	0.0036	0.0351	0.0519
PPV 95%	Random Forest	50,000	0.0046	stratified	7.8e-05 (0.00014)	8.3e-06 (5.2e-07)	8.3e-06	0.0038	0.0348	0.0519
PPV 95%	Random Forest	50,000	0.0092	naive	-0.00031 (0.0002)	1.7e-05 (1e-06)	1.7e-05	0.0055	0.0758	0.0977
PPV 95%	Random Forest	50,000	0.0092	stratified	-0.0002 (0.00021)	1.8e-05 (1.1e-06)	1.8e-05	0.0058	0.074	0.1
PPV 95%	Random Forest	50,000	0.0184	naive	-0.001 (0.00029)	3.2e-05 (2e-06)	3.1e-05	0.0069	0.155	0.187
PPV 95%	Random Forest	50,000	0.0184	stratified	-0.00053 (0.00029)	3.5e-05 (2.1e-06)	3.4e-05	0.0078	0.153	0.187
PPV 95%	Random Forest	100,000	0.0046	naive	-0.00013 (0.00011)	4.4e-06 (3e-07)	4.4e-06	0.0025	0.038	0.0507
PPV 95%	Random Forest	100,000	0.0046	stratified	-0.00017 (0.0001)	4.8e-06 (2.9e-07)	4.8e-06	0.003	0.0387	0.0518
PPV 95%	Random Forest	100,000	0.0092	naive	-8.2e-05 (0.00013)	7.1e-06 (4.3e-07)	7.2e-06	0.0038	0.0811	0.0965
PPV 95%	Random Forest	100,000	0.0092	stratified	1.5e-06 (0.00014)	7.8e-06 (4.8e-07)	7.8e-06	0.0039	0.0791	0.0972
PPV 95%	Random Forest	100,000	0.0184	naive	-0.00049 (0.00021)	1.7e-05 (1.1e-06)	1.7e-05	0.0055	0.159	0.187
PPV 95%	Random Forest	100,000	0.0184	stratified	-0.00019 (0.00021)	1.7e-05 (9.9e-07)	1.7e-05	0.0055	0.163	0.185
PPV 95%	Random Forest	1,000,000	0.0046	naive	-0.00013 (4.3e-05)	4e-07 (2.4e-08)	3.8e-07	0.00084	0.0435	0.0472
PPV 95%	Random Forest	1,000,000	0.0046	stratified	-0.00016 (4.1e-05)	3.4e-07 (1.9e-08)	3.1e-07	0.00075	0.0439	0.0471
PPV 95%	Random Forest	1,000,000	0.0092	naive	-0.00019 (6.3e-05)	6.3e-07 (4e-08)	5.9e-07	0.0011	0.0868	0.0914
PPV 95%	Random Forest	1,000,000	0.0092	stratified	-0.00026 (6e-05)	6.1e-07 (3.9e-08)	5.5e-07	0.001	0.0872	0.0915

Table C.44: Bias, MSE, and Variability for PPV at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 99%	Random Forest	5,000	0.0046	naive	0.0085 (0.0025)	0.0033 (0.00026)	0.0033	0.072	0	0.352
PPV 99%	Random Forest	5,000	0.0046	stratified	0.0045 (0.0024)	0.003 (0.0002)	0.003	0.079	0	0.298
PPV 99%	Random Forest	5,000	0.0092	naive	-0.0013 (0.003)	0.0047 (0.00031)	0.0047	0.092	0.01	0.442
PPV 99%	Random Forest	5,000	0.0092	stratified	0.0071 (0.0033)	0.0057 (0.00032)	0.0056	0.11	0.02	0.397
PPV 99%	Random Forest	5,000	0.0184	naive	-0.0028 (0.0039)	0.0075 (0.00048)	0.0075	0.12	0.0889	0.647
PPV 99%	Random Forest	5,000	0.0184	stratified	0.0083 (0.0038)	0.007 (0.00043)	0.0069	0.11	0.1	0.579
PPV 99%	Random Forest	10,000	0.0046	naive	0.0021 (0.0014)	0.001 (6e-05)	0.001	0.047	0.0143	0.183
PPV 99%	Random Forest	10,000	0.0046	stratified	-0.0011 (0.0014)	0.001 (6.5e-05)	0.001	0.044	0	0.199
PPV 99%	Random Forest	10,000	0.0092	naive	-0.0046 (0.0021)	0.0021 (0.00013)	0.0021	0.065	0.0551	0.328
PPV 99%	Random Forest	10,000	0.0092	stratified	0.0044 (0.0021)	0.0022 (0.00014)	0.0021	0.062	0.0614	0.397
PPV 99%	Random Forest	10,000	0.0184	naive	-0.0034 (0.0025)	0.0028 (0.00019)	0.0028	0.072	0.232	0.57
PPV 99%	Random Forest	10,000	0.0184	stratified	0.0061 (0.0026)	0.0032 (0.00019)	0.0032	0.079	0.232	0.55
PPV 99%	Random Forest	50,000	0.0046	naive	-0.00041 (0.00061)	0.00017 (1e-05)	0.00017	0.016	0.0657	0.142
PPV 99%	Random Forest	50,000	0.0046	stratified	6.5e-07 (0.0006)	0.00016 (9.8e-06)	0.00016	0.018	0.07	0.147
PPV 99%	Random Forest	50,000	0.0092	naive	-0.0002 (0.00077)	0.00025 (1.6e-05)	0.00025	0.021	0.154	0.255
PPV 99%	Random Forest	50,000	0.0092	stratified	-0.00021 (0.00077)	0.00025 (1.6e-05)	0.00025	0.02	0.164	0.27
PPV 99%	Random Forest	50,000	0.0184	naive	-0.0013 (0.00094)	0.00036 (2.3e-05)	0.00036	0.025	0.317	0.433
PPV 99%	Random Forest	50,000	0.0184	stratified	0.00019 (0.00096)	0.00037 (2.6e-05)	0.00037	0.025	0.311	0.447
PPV 99%	Random Forest	100,000	0.0046	naive	-0.00025 (0.00043)	7.7e-05 (4.6e-06)	7.7e-05	0.012	0.0834	0.136
PPV 99%	Random Forest	100,000	0.0046	stratified	-0.00067 (0.00042)	7.4e-05 (5.1e-06)	7.4e-05	0.011	0.0796	0.132
PPV 99%	Random Forest	100,000	0.0092	naive	-0.00057 (0.00054)	0.00012 (7.7e-06)	0.00012	0.015	0.174	0.243
PPV 99%	Random Forest	100,000	0.0092	stratified	0.00082 (0.00056)	0.00012 (7.9e-06)	0.00012	0.015	0.181	0.245
PPV 99%	Random Forest	100,000	0.0184	naive	-0.0011 (0.00068)	0.00017 (1.1e-05)	0.00017	0.018	0.344	0.425
PPV 99%	Random Forest	100,000	0.0184	stratified	-0.00091 (0.00068)	0.00019 (1.2e-05)	0.00019	0.019	0.33	0.41
PPV 99%	Random Forest	1,000,000	0.0046	naive	-0.00037 (0.00019)	6.1e-06 (4.2e-07)	6e-06	0.0031	0.1	0.116
PPV 99%	Random Forest	1,000,000	0.0046	stratified	-0.00014 (0.00017)	5e-06 (3e-07)	5e-06	0.003	0.102	0.114
PPV 99%	Random Forest	1,000,000	0.0092	naive	-0.00048 (0.00024)	8.5e-06 (4.9e-07)	8.3e-06	0.0042	0.198	0.213
PPV 99%	Random Forest	1,000,000	0.0092	stratified	-0.00041 (0.00024)	8.4e-06 (5.4e-07)	8.3e-06	0.0039	0.197	0.214

Table C.45: Bias, MSE, and Variability for PPV at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 90%	Ridge	5,000	0.0046	naive	0.0028 (0.00019)	4.4e-05 (2.1e-06)	3.7e-05	0.0082	0.0057	0.0499
PPV 90%	Ridge	5,000	0.0046	stratified	9.3e-05 (0.00016)	2.7e-05 (1.3e-06)	2.7e-05	0.0074	0.00182	0.0439
PPV 90%	Ridge	5,000	0.0092	naive	0.00062 (0.00023)	5.1e-05 (2.4e-06)	5.1e-05	0.01	0.0277	0.0759
PPV 90%	Ridge	5,000	0.0092	stratified	0.00021 (0.00023)	4.8e-05 (2.1e-06)	4.8e-05	0.0097	0.0318	0.0762
PPV 90%	Ridge	5,000	0.0184	naive	-2.1e-05 (0.00031)	9.1e-05 (4e-06)	9.1e-05	0.013	0.0746	0.133
PPV 90%	Ridge	5,000	0.0184	stratified	0.00051 (0.00031)	8.9e-05 (3.9e-06)	8.9e-05	0.013	0.0754	0.137
PPV 90%	Ridge	10,000	0.0046	naive	0.00028 (0.00012)	1.3e-05 (5.5e-07)	1.3e-05	0.0053	0.0152	0.0392
PPV 90%	Ridge	10,000	0.0046	stratified	-1.7e-05 (0.00011)	1.1e-05 (5.2e-07)	1.1e-05	0.0045	0.0149	0.0372
PPV 90%	Ridge	10,000	0.0092	naive	-0.00015 (0.00015)	2.2e-05 (9.7e-07)	2.2e-05	0.0066	0.0393	0.0686
PPV 90%	Ridge	10,000	0.0092	stratified	9e-05 (0.00016)	2.4e-05 (1e-06)	2.4e-05	0.0065	0.0393	0.0686
PPV 90%	Ridge	10,000	0.0184	naive	-0.00026 (0.00022)	4.5e-05 (2e-06)	4.5e-05	0.0092	0.0859	0.128
PPV 90%	Ridge	10,000	0.0184	stratified	0.00017 (0.00023)	4.6e-05 (2.1e-06)	4.6e-05	0.0091	0.0872	0.134
PPV 90%	Ridge	50,000	0.0046	naive	-6.9e-05 (5.1e-05)	2.4e-06 (1.1e-07)	2.4e-06	0.0022	0.0229	0.0333
PPV 90%	Ridge	50,000	0.0046	stratified	-7.8e-05 (4.8e-05)	2.1e-06 (9.9e-08)	2.1e-06	0.002	0.0224	0.0333
PPV 90%	Ridge	50,000	0.0092	naive	-2.4e-05 (6.8e-05)	4.2e-06 (1.8e-07)	4.2e-06	0.0027	0.0508	0.063
PPV 90%	Ridge	50,000	0.0092	stratified	5.9e-06 (6.9e-05)	4.2e-06 (1.9e-07)	4.2e-06	0.0028	0.0492	0.0632
PPV 90%	Ridge	50,000	0.0184	naive	-0.00014 (0.0001)	8.5e-06 (4e-07)	8.5e-06	0.0037	0.104	0.123
PPV 90%	Ridge	50,000	0.0184	stratified	6.3e-05 (0.0001)	8.5e-06 (3.8e-07)	8.6e-06	0.0039	0.103	0.123
PPV 90%	Ridge	100,000	0.0046	naive	-2.1e-05 (3.5e-05)	1.1e-06 (5e-08)	1.1e-06	0.0014	0.0254	0.0324
PPV 90%	Ridge	100,000	0.0046	stratified	5e-05 (3.7e-05)	1.2e-06 (5.5e-08)	1.2e-06	0.0015	0.0253	0.0319
PPV 90%	Ridge	100,000	0.0092	naive	-4.4e-05 (5e-05)	2.1e-06 (9.2e-08)	2.1e-06	0.002	0.0521	0.0627
PPV 90%	Ridge	100,000	0.0092	stratified	3.4e-05 (5e-05)	2.1e-06 (9.6e-08)	2.1e-06	0.0019	0.0531	0.0619
PPV 90%	Ridge	100,000	0.0184	naive	-6.8e-05 (7e-05)	3.8e-06 (1.7e-07)	3.8e-06	0.0028	0.108	0.122
PPV 90%	Ridge	100,000	0.0184	stratified	8.5e-05 (7.2e-05)	3.9e-06 (1.9e-07)	3.9e-06	0.0025	0.108	0.123
PPV 90%	Ridge	1,000,000	0.0046	naive	-6.8e-07 (1.5e-05)	9.2e-08 (3.9e-09)	9.2e-08	0.0004	0.0285	0.0305
PPV 90%	Ridge	1,000,000	0.0046	stratified	-1.6e-05 (1.5e-05)	9.3e-08 (3.8e-09)	9.3e-08	0.00045	0.0286	0.0304
PPV 90%	Ridge	1,000,000	0.0092	naive	-3.6e-05 (2.1e-05)	1.4e-07 (6.2e-09)	1.4e-07	0.00048	0.0576	0.0598
PPV 90%	Ridge	1,000,000	0.0092	stratified	-4.8e-05 (2e-05)	1.3e-07 (5.9e-09)	1.3e-07	0.00052	0.0574	0.0599

Table C.46: Bias, MSE, and Variability for PPV at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 95%	Ridge	5,000	0.0046	naive	0.0044 (0.00038)	0.00017 (8.1e-06)	0.00015	0.016	0.0109	0.0939
PPV 95%	Ridge	5,000	0.0046	stratified	-4.3e-05 (0.00033)	0.00011 (5e-06)	0.00011	0.014	0	0.0707
PPV 95%	Ridge	5,000	0.0092	naive	0.00073 (0.00046)	0.00021 (9.4e-06)	0.00021	0.02	0.033	0.133
PPV 95%	Ridge	5,000	0.0092	stratified	0.00082 (0.00046)	0.00021 (9.5e-06)	0.00021	0.02	0.0348	0.129
PPV 95%	Ridge	5,000	0.0184	naive	-3.9e-05 (0.00063)	0.00038 (1.7e-05)	0.00038	0.026	0.0884	0.221
PPV 95%	Ridge	5,000	0.0184	stratified	0.0019 (0.00063)	0.00037 (1.6e-05)	0.00037	0.026	0.101	0.217
PPV 95%	Ridge	10,000	0.0046	naive	0.00042 (0.00023)	5.2e-05 (2.3e-06)	5.2e-05	0.01	0.0155	0.0631
PPV 95%	Ridge	10,000	0.0046	stratified	9e-05 (0.00021)	4.6e-05 (2.1e-06)	4.6e-05	0.0094	0.0223	0.0634
PPV 95%	Ridge	10,000	0.0092	naive	-0.00021 (0.00031)	9.2e-05 (4.1e-06)	9.2e-05	0.013	0.0552	0.112
PPV 95%	Ridge	10,000	0.0092	stratified	-1.9e-05 (0.00032)	9.9e-05 (4.2e-06)	9.9e-05	0.013	0.0537	0.113
PPV 95%	Ridge	10,000	0.0184	naive	-0.00022 (0.00046)	0.00019 (8.4e-06)	0.00019	0.019	0.123	0.207
PPV 95%	Ridge	10,000	0.0184	stratified	0.00074 (0.00044)	0.00018 (7.8e-06)	0.00018	0.018	0.126	0.206
PPV 95%	Ridge	50,000	0.0046	naive	-0.00015 (0.0001)	9.8e-06 (4.5e-07)	9.8e-06	0.0042	0.0314	0.0543
PPV 95%	Ridge	50,000	0.0046	stratified	-0.00015 (9.9e-05)	9e-06 (4.1e-07)	9e-06	0.0043	0.0338	0.0542
PPV 95%	Ridge	50,000	0.0092	naive	-0.00017 (0.00014)	1.7e-05 (7.8e-07)	1.7e-05	0.0055	0.073	0.1
PPV 95%	Ridge	50,000	0.0092	stratified	0.0001 (0.00014)	1.7e-05 (7.6e-07)	1.7e-05	0.0056	0.0749	0.101
PPV 95%	Ridge	50,000	0.0184	naive	-0.00032 (0.0002)	3.5e-05 (1.7e-06)	3.5e-05	0.0074	0.153	0.198
PPV 95%	Ridge	50,000	0.0184	stratified	6.4e-05 (0.00019)	3.3e-05 (1.5e-06)	3.3e-05	0.0077	0.157	0.198
PPV 95%	Ridge	100,000	0.0046	naive	-4.6e-05 (7.3e-05)	4.8e-06 (2.1e-07)	4.8e-06	0.003	0.0376	0.0522
PPV 95%	Ridge	100,000	0.0046	stratified	4.9e-05 (7.4e-05)	4.9e-06 (2.1e-07)	4.9e-06	0.0029	0.0379	0.0516
PPV 95%	Ridge	100,000	0.0092	naive	-0.00013 (0.0001)	8.8e-06 (3.9e-07)	8.7e-06	0.0041	0.0813	0.1
PPV 95%	Ridge	100,000	0.0092	stratified	-5.9e-05 (1e-04)	8.7e-06 (3.9e-07)	8.7e-06	0.0038	0.0812	0.0992
PPV 95%	Ridge	100,000	0.0184	naive	-9.1e-05 (0.00014)	1.6e-05 (6.8e-07)	1.6e-05	0.0054	0.163	0.189
PPV 95%	Ridge	100,000	0.0184	stratified	6.1e-05 (0.00014)	1.7e-05 (8.5e-07)	1.7e-05	0.0054	0.163	0.196
PPV 95%	Ridge	1,000,000	0.0046	naive	-2.3e-06 (3.1e-05)	3.8e-07 (1.7e-08)	3.8e-07	0.00082	0.045	0.0487
PPV 95%	Ridge	1,000,000	0.0046	stratified	-2.9e-05 (3e-05)	4.1e-07 (1.8e-08)	4.1e-07	0.00086	0.0449	0.0488
PPV 95%	Ridge	1,000,000	0.0092	naive	-6.2e-05 (4.2e-05)	5.8e-07 (2.5e-08)	5.8e-07	0.001	0.0902	0.0949
PPV 95%	Ridge	1,000,000	0.0092	stratified	-3.7e-05 (4.1e-05)	5.7e-07 (2.6e-08)	5.7e-07	0.00098	0.0896	0.0948

Table C.47: Bias, MSE, and Variability for PPV at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 99%	Ridge	5,000	0.0046	naive	0.012 (0.0017)	0.0033 (0.00017)	0.0031	0.076	0	0.325
PPV 99%	Ridge	5,000	0.0046	stratified	-0.0008 (0.0015)	0.0025 (0.00013)	0.0025	0.068	0	0.352
PPV 99%	Ridge	5,000	0.0092	naive	-0.0014 (0.002)	0.0041 (0.00018)	0.0041	0.087	0.0167	0.396
PPV 99%	Ridge	5,000	0.0092	stratified	0.0044 (0.002)	0.0045 (0.0002)	0.0045	0.092	0.0167	0.42
PPV 99%	Ridge	5,000	0.0184	naive	0.00062 (0.0024)	0.0062 (0.00028)	0.0062	0.11	0.124	0.652
PPV 99%	Ridge	5,000	0.0184	stratified	0.0071 (0.0025)	0.0064 (0.00028)	0.0064	0.11	0.116	0.61
PPV 99%	Ridge	10,000	0.0046	naive	0.0018 (0.00096)	0.00096 (4.3e-05)	0.00096	0.04	0.02	0.202
PPV 99%	Ridge	10,000	0.0046	stratified	0.0015 (0.001)	0.0011 (5.4e-05)	0.0011	0.044	0.0154	0.234
PPV 99%	Ridge	10,000	0.0092	naive	-0.00044 (0.0013)	0.0019 (8.3e-05)	0.0019	0.06	0.0774	0.375
PPV 99%	Ridge	10,000	0.0092	stratified	0.0011 (0.0014)	0.0019 (8.8e-05)	0.0019	0.059	0.0875	0.346
PPV 99%	Ridge	10,000	0.0184	naive	0.0028 (0.0017)	0.0029 (0.00013)	0.0029	0.071	0.193	0.57
PPV 99%	Ridge	10,000	0.0184	stratified	0.002 (0.0016)	0.0028 (0.00011)	0.0028	0.077	0.237	0.533
PPV 99%	Ridge	50,000	0.0046	naive	-0.00072 (0.0004)	0.00016 (7.3e-06)	0.00016	0.016	0.0671	0.153
PPV 99%	Ridge	50,000	0.0046	stratified	-0.00089 (0.00042)	0.00017 (7e-06)	0.00017	0.018	0.0696	0.146
PPV 99%	Ridge	50,000	0.0092	naive	-0.00037 (0.00055)	0.00029 (1.3e-05)	0.00029	0.022	0.154	0.262
PPV 99%	Ridge	50,000	0.0092	stratified	0.00056 (0.00057)	0.00031 (1.4e-05)	0.00031	0.024	0.161	0.263
PPV 99%	Ridge	50,000	0.0184	naive	-6e-05 (0.00067)	0.00044 (2e-05)	0.00044	0.03	0.323	0.454
PPV 99%	Ridge	50,000	0.0184	stratified	-0.00095 (0.00067)	0.00044 (2.1e-05)	0.00044	0.027	0.308	0.452
PPV 99%	Ridge	100,000	0.0046	naive	-0.00024 (0.00029)	8e-05 (3.4e-06)	8e-05	0.012	0.0844	0.139
PPV 99%	Ridge	100,000	0.0046	stratified	0.0003 (0.0003)	8.3e-05 (3.5e-06)	8.3e-05	0.012	0.0882	0.139
PPV 99%	Ridge	100,000	0.0092	naive	-0.00094 (0.0004)	0.00015 (6.4e-06)	0.00015	0.016	0.18	0.253
PPV 99%	Ridge	100,000	0.0092	stratified	-0.0002 (0.00039)	0.00015 (6.5e-06)	0.00015	0.016	0.177	0.255
PPV 99%	Ridge	100,000	0.0184	naive	-0.0008 (0.00049)	0.00022 (1e-05)	0.00022	0.02	0.347	0.444
PPV 99%	Ridge	100,000	0.0184	stratified	0.00042 (0.0005)	0.00022 (1e-05)	0.00022	0.02	0.337	0.438
PPV 99%	Ridge	1,000,000	0.0046	naive	-4.9e-05 (0.00013)	7.2e-06 (3.4e-07)	7.2e-06	0.0036	0.108	0.126
PPV 99%	Ridge	1,000,000	0.0046	stratified	-6.6e-05 (0.00013)	7.1e-06 (3.1e-07)	7.1e-06	0.0036	0.108	0.126
PPV 99%	Ridge	1,000,000	0.0092	naive	-0.00025 (0.00016)	9.3e-06 (4.2e-07)	9.2e-06	0.004	0.214	0.234
PPV 99%	Ridge	1,000,000	0.0092	stratified	-0.00033 (0.00016)	9.7e-06 (4.3e-07)	9.6e-06	0.0042	0.214	0.234

Table C.48: Bias, MSE, and Variability for PPV at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 90%	Logistic	5,000	0.0046	naive	-0.00043 (7.9e-05)	4.2e-06 (3.1e-07)	4e-06	0.0016	0	0.0171
PPV 90%	Logistic	5,000	0.0046	stratified	-0.00073 (7.5e-05)	4.1e-06 (2.4e-07)	3.6e-06	0.0011	0	0.0174
PPV 90%	Logistic	5,000	0.0092	naive	-0.0029 (0.00023)	5.7e-05 (2.3e-06)	4.9e-05	0.0093	0.00405	0.0464
PPV 90%	Logistic	5,000	0.0092	stratified	-0.0019 (0.00023)	5.5e-05 (2.3e-06)	5.1e-05	0.0099	0.00411	0.0441
PPV 90%	Logistic	5,000	0.0184	naive	-0.0035 (0.00038)	0.00016 (1e-05)	0.00014	0.015	0.0175	0.107
PPV 90%	Logistic	5,000	0.0184	stratified	-0.0035 (0.00039)	0.00017 (1.2e-05)	0.00016	0.015	0.0177	0.111
PPV 90%	Logistic	10,000	0.0046	naive	-0.001 (0.00012)	1.3e-05 (5.7e-07)	1.2e-05	0.0048	0.00248	0.0246
PPV 90%	Logistic	10,000	0.0046	stratified	-0.00082 (0.00012)	1.2e-05 (5.2e-07)	1.2e-05	0.0047	0.00351	0.0225
PPV 90%	Logistic	10,000	0.0092	naive	-0.0019 (0.00019)	3.4e-05 (1.5e-06)	3.1e-05	0.0075	0.02	0.0552
PPV 90%	Logistic	10,000	0.0092	stratified	-0.0018 (0.00018)	3.1e-05 (1.4e-06)	2.7e-05	0.0071	0.0187	0.0538
PPV 90%	Logistic	10,000	0.0184	naive	-0.0021 (0.00024)	5.8e-05 (2.5e-06)	5.4e-05	0.01	0.0719	0.113
PPV 90%	Logistic	10,000	0.0184	stratified	-0.0025 (0.00025)	6.1e-05 (2.9e-06)	5.4e-05	0.01	0.0635	0.115
PPV 90%	Logistic	50,000	0.0046	naive	-0.00042 (5.3e-05)	2.7e-06 (1.2e-07)	2.5e-06	0.002	0.0202	0.03
PPV 90%	Logistic	50,000	0.0046	stratified	-0.00039 (5.4e-05)	2.7e-06 (1.2e-07)	2.6e-06	0.0021	0.02	0.0307
PPV 90%	Logistic	50,000	0.0092	naive	-0.00036 (7.4e-05)	5.2e-06 (2.4e-07)	5.1e-06	0.003	0.0481	0.0618
PPV 90%	Logistic	50,000	0.0092	stratified	-0.00019 (7e-05)	4.5e-06 (2e-07)	4.5e-06	0.0029	0.048	0.0625
PPV 90%	Logistic	50,000	0.0184	naive	-0.00037 (0.0001)	9.3e-06 (4.3e-07)	9.2e-06	0.0042	0.101	0.121
PPV 90%	Logistic	50,000	0.0184	stratified	-0.00028 (0.0001)	9.2e-06 (4e-07)	9.2e-06	0.0042	0.103	0.122
PPV 90%	Logistic	100,000	0.0046	naive	-0.00014 (3.8e-05)	1.3e-06 (6.3e-08)	1.3e-06	0.0015	0.0236	0.0317
PPV 90%	Logistic	100,000	0.0046	stratified	-0.00015 (3.7e-05)	1.2e-06 (5.5e-08)	1.2e-06	0.0015	0.0241	0.0308
PPV 90%	Logistic	100,000	0.0092	naive	-0.00017 (5.1e-05)	2.2e-06 (1e-07)	2.2e-06	0.002	0.052	0.0609
PPV 90%	Logistic	100,000	0.0092	stratified	-0.00014 (5.2e-05)	2.3e-06 (1.1e-07)	2.3e-06	0.0021	0.0503	0.0612
PPV 90%	Logistic	100,000	0.0184	naive	-0.00011 (7.2e-05)	4.1e-06 (1.8e-07)	4.1e-06	0.0028	0.108	0.12
PPV 90%	Logistic	100,000	0.0184	stratified	-0.00017 (7e-05)	3.9e-06 (1.7e-07)	3.9e-06	0.0028	0.108	0.12
PPV 90%	Logistic	1,000,000	0.0046	naive	-1.9e-05 (1.5e-05)	9e-08 (4e-09)	9e-08	0.00042	0.0285	0.0304
PPV 90%	Logistic	1,000,000	0.0046	stratified	-5.5e-06 (1.5e-05)	8.7e-08 (3.7e-09)	8.7e-08	0.00041	0.0286	0.0303
PPV 90%	Logistic	1,000,000	0.0092	naive	-8.3e-07 (2.1e-05)	1.4e-07 (6.6e-09)	1.4e-07	0.00051	0.0574	0.0599
PPV 90%	Logistic	1,000,000	0.0092	stratified	2.6e-06 (2.1e-05)	1.4e-07 (6.3e-09)	1.4e-07	0.00053	0.0572	0.0599

Table C.49: Bias, MSE, and Variability for PPV at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 95%	Logistic	5,000	0.0046	naive	-0.0012 (0.00014)	1.5e-05 (9.8e-07)	1.3e-05	0.0028	0	0.0278
PPV 95%	Logistic	5,000	0.0046	stratified	-0.0015 (0.00013)	1.4e-05 (6.8e-07)	1.2e-05	0.0026	0	0.0243
PPV 95%	Logistic	5,000	0.0092	naive	-0.0045 (0.0004)	0.00018 (7.4e-06)	0.00016	0.017	0.00385	0.0784
PPV 95%	Logistic	5,000	0.0092	stratified	-0.0029 (0.00041)	0.00017 (7e-06)	0.00016	0.017	0	0.0719
PPV 95%	Logistic	5,000	0.0184	naive	-0.0054 (0.00067)	0.00049 (2.8e-05)	0.00047	0.028	0.0175	0.17
PPV 95%	Logistic	5,000	0.0184	stratified	-0.005 (0.0007)	0.00055 (3.2e-05)	0.00052	0.029	0.0178	0.171
PPV 95%	Logistic	10,000	0.0046	naive	-0.0019 (0.00021)	4.1e-05 (1.8e-06)	3.8e-05	0.0081	0	0.0378
PPV 95%	Logistic	10,000	0.0046	stratified	-0.0013 (0.0002)	4e-05 (1.6e-06)	3.8e-05	0.0085	0.00311	0.0367
PPV 95%	Logistic	10,000	0.0092	naive	-0.0033 (0.00034)	0.00012 (5e-06)	0.00011	0.014	0.0201	0.0935
PPV 95%	Logistic	10,000	0.0092	stratified	-0.003 (0.00033)	0.00011 (4.9e-06)	0.0001	0.014	0.0267	0.0984
PPV 95%	Logistic	10,000	0.0184	naive	-0.0038 (0.00046)	0.00022 (9.2e-06)	0.0002	0.019	0.0975	0.183
PPV 95%	Logistic	10,000	0.0184	stratified	-0.0037 (0.00047)	0.00022 (9.5e-06)	0.0002	0.019	0.0947	0.192
PPV 95%	Logistic	50,000	0.0046	naive	-0.00081 (0.0001)	1.1e-05 (4.6e-07)	1e-05	0.0042	0.0288	0.0481
PPV 95%	Logistic	50,000	0.0046	stratified	-0.00077 (0.00011)	1.1e-05 (4.9e-07)	1e-05	0.0042	0.0286	0.0497
PPV 95%	Logistic	50,000	0.0092	naive	-0.00072 (0.00014)	2e-05 (9.9e-07)	2e-05	0.0057	0.07	0.1
PPV 95%	Logistic	50,000	0.0092	stratified	-0.0004 (0.00014)	1.9e-05 (8.5e-07)	1.9e-05	0.0058	0.0713	0.0995
PPV 95%	Logistic	50,000	0.0184	naive	-0.00079 (0.00019)	3.6e-05 (1.5e-06)	3.5e-05	0.0083	0.155	0.192
PPV 95%	Logistic	50,000	0.0184	stratified	-0.00039 (0.0002)	3.8e-05 (1.7e-06)	3.7e-05	0.0082	0.153	0.196
PPV 95%	Logistic	100,000	0.0046	naive	-0.00031 (7.7e-05)	5.4e-06 (2.4e-07)	5.4e-06	0.0031	0.0367	0.0514
PPV 95%	Logistic	100,000	0.0046	stratified	-0.00029 (7.5e-05)	5.2e-06 (2.3e-07)	5.1e-06	0.0031	0.0363	0.0517
PPV 95%	Logistic	100,000	0.0092	naive	-0.00037 (0.00011)	9.9e-06 (4.2e-07)	9.8e-06	0.0043	0.0792	0.0983
PPV 95%	Logistic	100,000	0.0092	stratified	-0.00031 (0.00011)	9.6e-06 (4.4e-07)	9.6e-06	0.0039	0.0798	0.0987
PPV 95%	Logistic	100,000	0.0184	naive	-0.00013 (0.00014)	1.7e-05 (7.2e-07)	1.7e-05	0.0059	0.165	0.19
PPV 95%	Logistic	100,000	0.0184	stratified	-0.00025 (0.00014)	1.6e-05 (6.6e-07)	1.6e-05	0.0054	0.163	0.186
PPV 95%	Logistic	1,000,000	0.0046	naive	-2.1e-05 (3.1e-05)	3.9e-07 (1.8e-08)	3.9e-07	0.00087	0.0448	0.0491
PPV 95%	Logistic	1,000,000	0.0046	stratified	-2.9e-06 (3.1e-05)	4e-07 (1.8e-08)	4e-07	0.00081	0.045	0.0487
PPV 95%	Logistic	1,000,000	0.0092	naive	9.5e-06 (4.2e-05)	6.2e-07 (3e-08)	6.2e-07	0.001	0.0902	0.0956
PPV 95%	Logistic	1,000,000	0.0092	stratified	1.1e-05 (4.3e-05)	6.3e-07 (2.8e-08)	6.3e-07	0.0011	0.0899	0.0954

Table C.50: Bias, MSE, and Variability for PPV at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
PPV 99%	Logistic	5,000	0.0046	naive	0.0011 (0.00051)	0.00024 (1.6e-05)	0.00024	0.019	0	0.11
PPV 99%	Logistic	5,000	0.0046	stratified	-0.00057 (0.00046)	0.0002 (1e-05)	0.0002	0.021	0	0.077
PPV 99%	Logistic	5,000	0.0092	naive	-0.01 (0.0011)	0.0013 (5.5e-05)	0.0012	0.045	0	0.216
PPV 99%	Logistic	5,000	0.0092	stratified	-0.0069 (0.0011)	0.0013 (6.1e-05)	0.0012	0.048	0	0.247
PPV 99%	Logistic	5,000	0.0184	naive	-0.012 (0.0019)	0.0039 (0.00018)	0.0038	0.082	0.0368	0.423
PPV 99%	Logistic	5,000	0.0184	stratified	-0.012 (0.002)	0.0043 (0.00019)	0.0041	0.085	0.0126	0.415
PPV 99%	Logistic	10,000	0.0046	naive	-0.0054 (0.00064)	0.00042 (1.6e-05)	0.00039	0.026	0	0.109
PPV 99%	Logistic	10,000	0.0046	stratified	-0.0032 (0.00068)	0.00046 (2.6e-05)	0.00046	0.028	0	0.169
PPV 99%	Logistic	10,000	0.0092	naive	-0.0089 (0.001)	0.0012 (5e-05)	0.0011	0.045	0.0291	0.241
PPV 99%	Logistic	10,000	0.0092	stratified	-0.007 (0.0011)	0.0013 (5.7e-05)	0.0013	0.048	0.0111	0.258
PPV 99%	Logistic	10,000	0.0184	naive	-0.0087 (0.0015)	0.0025 (0.00011)	0.0024	0.064	0.146	0.464
PPV 99%	Logistic	10,000	0.0184	stratified	-0.01 (0.0016)	0.0026 (0.0001)	0.0025	0.069	0.143	0.453
PPV 99%	Logistic	50,000	0.0046	naive	-0.0022 (0.0004)	0.00017 (6.8e-06)	0.00016	0.018	0.054	0.134
PPV 99%	Logistic	50,000	0.0046	stratified	-0.0024 (0.0004)	0.00017 (7.5e-06)	0.00017	0.018	0.055	0.142
PPV 99%	Logistic	50,000	0.0092	naive	-0.0024 (0.00056)	0.00032 (1.4e-05)	0.00031	0.024	0.146	0.257
PPV 99%	Logistic	50,000	0.0092	stratified	-0.0017 (0.00055)	0.00031 (1.3e-05)	0.0003	0.024	0.145	0.256
PPV 99%	Logistic	50,000	0.0184	naive	-0.0018 (0.00068)	0.00047 (2.1e-05)	0.00047	0.029	0.295	0.449
PPV 99%	Logistic	50,000	0.0184	stratified	-0.0011 (0.00072)	0.0005 (2.2e-05)	0.0005	0.031	0.309	0.455
PPV 99%	Logistic	100,000	0.0046	naive	-0.0011 (0.0003)	8.7e-05 (3.9e-06)	8.6e-05	0.013	0.073	0.138
PPV 99%	Logistic	100,000	0.0046	stratified	-0.00081 (0.0003)	8.9e-05 (3.9e-06)	8.8e-05	0.012	0.0794	0.139
PPV 99%	Logistic	100,000	0.0092	naive	-0.0015 (0.00041)	0.00016 (7e-06)	0.00016	0.017	0.175	0.251
PPV 99%	Logistic	100,000	0.0092	stratified	-0.0016 (0.00041)	0.00016 (7.2e-06)	0.00015	0.016	0.167	0.253
PPV 99%	Logistic	100,000	0.0184	naive	-0.00029 (0.00049)	0.00022 (1e-05)	0.00022	0.02	0.331	0.444
PPV 99%	Logistic	100,000	0.0184	stratified	-0.00072 (0.00048)	0.00021 (9.6e-06)	0.00021	0.019	0.342	0.434
PPV 99%	Logistic	1,000,000	0.0046	naive	-9.2e-05 (0.00013)	7.4e-06 (3.2e-07)	7.4e-06	0.0036	0.109	0.126
PPV 99%	Logistic	1,000,000	0.0046	stratified	-0.00023 (0.00013)	7.2e-06 (3.2e-07)	7.2e-06	0.0036	0.11	0.127
PPV 99%	Logistic	1,000,000	0.0092	naive	3.7e-05 (0.00017)	1.1e-05 (5e-07)	1.1e-05	0.0046	0.215	0.236
PPV 99%	Logistic	1,000,000	0.0092	stratified	0.00018 (0.00017)	1.1e-05 (4.7e-07)	1.1e-05	0.0043	0.213	0.233

Table C.51: Bias, MSE, and Variability for PPV at the 99th percentile, logistic regression.

C.8 F_1 *Score*

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 90%	Random Forest	5,000	0.0046	naive	0.0039 (0.00051)	0.00015 (1.2e-05)	0.00014	0.015	0.0205	0.105
F1 90%	Random Forest	5,000	0.0046	stratified	-0.00017 (0.00042)	9.5e-05 (5.9e-06)	9.5e-05	0.013	0.0172	0.0742
F1 90%	Random Forest	5,000	0.0092	naive	-0.00059 (0.00058)	0.00017 (9.8e-06)	0.00017	0.018	0.0558	0.123
F1 90%	Random Forest	5,000	0.0092	stratified	-0.00025 (0.00055)	0.00016 (1e-05)	0.00016	0.017	0.0566	0.131
F1 90%	Random Forest	5,000	0.0184	naive	-0.0027 (0.00072)	0.00024 (1.4e-05)	0.00023	0.022	0.133	0.219
F1 90%	Random Forest	5,000	0.0184	stratified	0.00097 (0.00072)	0.00024 (1.5e-05)	0.00024	0.022	0.127	0.224
F1 90%	Random Forest	10,000	0.0046	naive	0.00042 (0.00032)	4.6e-05 (2.6e-06)	4.5e-05	0.0098	0.032	0.0677
F1 90%	Random Forest	10,000	0.0046	stratified	0.00029 (0.00031)	4.1e-05 (2.6e-06)	4.1e-05	0.0086	0.0296	0.0684
F1 90%	Random Forest	10,000	0.0092	naive	-0.00086 (0.0004)	8.1e-05 (5.2e-06)	8e-05	0.012	0.0693	0.12
F1 90%	Random Forest	10,000	0.0092	stratified	0.0006 (0.00038)	7e-05 (4.2e-06)	7e-05	0.011	0.0734	0.121
F1 90%	Random Forest	10,000	0.0184	naive	-0.0021 (0.00053)	0.00014 (8e-06)	0.00013	0.016	0.147	0.214
F1 90%	Random Forest	10,000	0.0184	stratified	0.0015 (0.0005)	0.00012 (7.6e-06)	0.00012	0.014	0.153	0.213
F1 90%	Random Forest	50,000	0.0046	naive	-0.00013 (0.00013)	7.5e-06 (4.6e-07)	7.5e-06	0.0038	0.0444	0.0602
F1 90%	Random Forest	50,000	0.0046	stratified	-5.2e-05 (0.00013)	7.3e-06 (4.5e-07)	7.3e-06	0.0037	0.0438	0.0611
F1 90%	Random Forest	50,000	0.0092	naive	-0.00042 (0.00018)	1.3e-05 (8.4e-07)	1.3e-05	0.0051	0.091	0.114
F1 90%	Random Forest	50,000	0.0092	stratified	-0.00013 (0.00018)	1.4e-05 (8.9e-07)	1.4e-05	0.0049	0.092	0.115
F1 90%	Random Forest	50,000	0.0184	naive	-0.001 (0.00024)	2.4e-05 (1.5e-06)	2.3e-05	0.0064	0.171	0.201
F1 90%	Random Forest	50,000	0.0184	stratified	-0.00036 (0.00024)	2.3e-05 (1.4e-06)	2.3e-05	0.0061	0.174	0.202
F1 90%	Random Forest	100,000	0.0046	naive	-0.00014 (9.5e-05)	3.7e-06 (2.4e-07)	3.7e-06	0.0025	0.0482	0.0606
F1 90%	Random Forest	100,000	0.0046	stratified	-4e-05 (9.6e-05)	4e-06 (2.5e-07)	4e-06	0.0026	0.0492	0.0607
F1 90%	Random Forest	100,000	0.0092	naive	-0.00015 (0.00012)	6.1e-06 (3.6e-07)	6.1e-06	0.0034	0.0965	0.111
F1 90%	Random Forest	100,000	0.0092	stratified	5.4e-06 (0.00013)	6.9e-06 (4.5e-07)	6.9e-06	0.0035	0.0972	0.114
F1 90%	Random Forest	100,000	0.0184	naive	-0.00049 (0.00017)	1.1e-05 (7e-07)	1.1e-05	0.0044	0.18	0.199
F1 90%	Random Forest	100,000	0.0184	stratified	-0.00048 (0.00017)	1.1e-05 (6.8e-07)	1.1e-05	0.0045	0.181	0.201
F1 90%	Random Forest	1,000,000	0.0046	naive	-0.00012 (4e-05)	3.3e-07 (2e-08)	3.1e-07	0.00076	0.0537	0.0571
F1 90%	Random Forest	1,000,000	0.0046	stratified	-0.00012 (4e-05)	3.1e-07 (2e-08)	3e-07	0.00074	0.0534	0.0572
F1 90%	Random Forest	1,000,000	0.0092	naive	-0.00025 (5.4e-05)	5.1e-07 (3.1e-08)	4.5e-07	0.00092	0.103	0.107
F1 90%	Random Forest	1,000,000	0.0092	stratified	-0.00029 (5.6e-05)	5e-07 (2.9e-08)	4.2e-07	0.00084	0.103	0.107

Table C.52: Bias, MSE, and Variability for F_1 score at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 95%	Random Forest	5,000	0.0046	naive	0.0037 (0.00094)	0.00048 (3.7e-05)	0.00046	0.026	0.0198	0.162
F1 95%	Random Forest	5,000	0.0046	stratified	-0.001 (0.00081)	0.00036 (2.2e-05)	0.00036	0.027	0.0069	0.123
F1 95%	Random Forest	5,000	0.0092	naive	-0.003 (0.0011)	0.00056 (3.8e-05)	0.00056	0.03	0.0531	0.196
F1 95%	Random Forest	5,000	0.0092	stratified	-0.00023 (0.001)	0.00054 (3.6e-05)	0.00054	0.03	0.0474	0.208
F1 95%	Random Forest	5,000	0.0184	naive	-0.0051 (0.0012)	0.0007 (4.2e-05)	0.00067	0.035	0.156	0.31
F1 95%	Random Forest	5,000	0.0184	stratified	-8.1e-05 (0.0012)	0.00071 (4.2e-05)	0.00071	0.036	0.156	0.302
F1 95%	Random Forest	10,000	0.0046	naive	0.00067 (0.00059)	0.00016 (9.6e-06)	0.00016	0.017	0.0347	0.107
F1 95%	Random Forest	10,000	0.0046	stratified	-8.7e-05 (0.0006)	0.00016 (9.1e-06)	0.00016	0.018	0.0359	0.107
F1 95%	Random Forest	10,000	0.0092	naive	-0.0028 (0.00074)	0.00028 (1.9e-05)	0.00027	0.022	0.0885	0.196
F1 95%	Random Forest	10,000	0.0092	stratified	0.0011 (0.0007)	0.00024 (1.6e-05)	0.00024	0.02	0.0908	0.186
F1 95%	Random Forest	10,000	0.0184	naive	-0.004 (0.00081)	0.00033 (2.1e-05)	0.00031	0.022	0.188	0.296
F1 95%	Random Forest	10,000	0.0184	stratified	-6.2e-06 (0.00083)	0.00032 (2e-05)	0.00033	0.023	0.199	0.3
F1 95%	Random Forest	50,000	0.0046	naive	-0.00037 (0.00025)	2.6e-05 (1.6e-06)	2.6e-05	0.0066	0.0641	0.0946
F1 95%	Random Forest	50,000	0.0046	stratified	9.8e-05 (0.00025)	2.8e-05 (1.8e-06)	2.8e-05	0.0069	0.0639	0.0952
F1 95%	Random Forest	50,000	0.0092	naive	-0.00092 (0.00033)	4.8e-05 (2.9e-06)	4.7e-05	0.0092	0.128	0.165
F1 95%	Random Forest	50,000	0.0092	stratified	-0.00048 (0.00035)	5.1e-05 (3.3e-06)	5.1e-05	0.0097	0.125	0.169
F1 95%	Random Forest	50,000	0.0184	naive	-0.0021 (0.0004)	7e-05 (4.4e-06)	6.6e-05	0.01	0.226	0.273
F1 95%	Random Forest	50,000	0.0184	stratified	-0.0012 (0.00042)	7.4e-05 (4.6e-06)	7.3e-05	0.011	0.223	0.273
F1 95%	Random Forest	100,000	0.0046	naive	-0.00038 (0.00019)	1.5e-05 (1e-06)	1.5e-05	0.0045	0.0696	0.0927
F1 95%	Random Forest	100,000	0.0046	stratified	-0.00034 (0.00019)	1.6e-05 (9.7e-07)	1.6e-05	0.0055	0.0709	0.0949
F1 95%	Random Forest	100,000	0.0092	naive	-0.00037 (0.00022)	2e-05 (1.2e-06)	2e-05	0.0064	0.137	0.163
F1 95%	Random Forest	100,000	0.0092	stratified	-8.2e-05 (0.00023)	2.2e-05 (1.4e-06)	2.2e-05	0.0066	0.134	0.164
F1 95%	Random Forest	100,000	0.0184	naive	-0.001 (0.00029)	3.6e-05 (2.4e-06)	3.5e-05	0.008	0.233	0.274
F1 95%	Random Forest	100,000	0.0184	stratified	-0.00052 (0.0003)	3.6e-05 (2.1e-06)	3.6e-05	0.0081	0.238	0.271
F1 95%	Random Forest	1,000,000	0.0046	naive	-0.00025 (7.9e-05)	1.3e-06 (8.1e-08)	1.3e-06	0.0015	0.0797	0.0863
F1 95%	Random Forest	1,000,000	0.0046	stratified	-0.0003 (7.4e-05)	1.1e-06 (6.5e-08)	1e-06	0.0014	0.0804	0.0863
F1 95%	Random Forest	1,000,000	0.0092	naive	-0.00033 (0.00011)	1.8e-06 (1.1e-07)	1.7e-06	0.0018	0.147	0.154
F1 95%	Random Forest	1,000,000	0.0092	stratified	-0.00043 (0.0001)	1.7e-06 (1.1e-07)	1.6e-06	0.0017	0.147	0.155

Table C.53: Bias, MSE, and Variability for F_1 score at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 99%	Random Forest	5,000	0.0046	naive	-0.0051 (0.0027)	0.004 (0.00026)	0.004	0.086	0	0.348
F1 99%	Random Forest	5,000	0.0046	stratified	-0.0036 (0.0026)	0.0039 (0.00021)	0.0039	0.089	0	0.31
F1 99%	Random Forest	5,000	0.0092	naive	-0.016 (0.0025)	0.0036 (0.00023)	0.0034	0.077	0.0133	0.405
F1 99%	Random Forest	5,000	0.0092	stratified	-0.0066 (0.0027)	0.004 (0.00022)	0.004	0.094	0.0222	0.34
F1 99%	Random Forest	5,000	0.0184	naive	-0.013 (0.0023)	0.0028 (0.00018)	0.0026	0.065	0.0551	0.38
F1 99%	Random Forest	5,000	0.0184	stratified	-0.008 (0.0022)	0.0024 (0.00016)	0.0023	0.062	0.0608	0.364
F1 99%	Random Forest	10,000	0.0046	naive	-0.0047 (0.0018)	0.0017 (9.2e-05)	0.0016	0.055	0.0182	0.23
F1 99%	Random Forest	10,000	0.0046	stratified	-0.0051 (0.0018)	0.0017 (0.00011)	0.0017	0.057	0	0.261
F1 99%	Random Forest	10,000	0.0092	naive	-0.012 (0.0018)	0.0019 (0.00012)	0.0018	0.059	0.0572	0.313
F1 99%	Random Forest	10,000	0.0092	stratified	-0.0031 (0.0019)	0.0018 (0.00011)	0.0018	0.059	0.0614	0.322
F1 99%	Random Forest	10,000	0.0184	naive	-0.01 (0.0016)	0.0013 (9.1e-05)	0.0012	0.045	0.134	0.376
F1 99%	Random Forest	10,000	0.0184	stratified	-0.0063 (0.0016)	0.0013 (8.4e-05)	0.0012	0.049	0.118	0.348
F1 99%	Random Forest	50,000	0.0046	naive	-0.0022 (0.00081)	0.0003 (1.9e-05)	0.0003	0.022	0.0881	0.191
F1 99%	Random Forest	50,000	0.0046	stratified	-0.0009 (0.0008)	0.00029 (1.8e-05)	0.00029	0.024	0.0942	0.201
F1 99%	Random Forest	50,000	0.0092	naive	-0.0021 (0.00078)	0.00027 (1.7e-05)	0.00026	0.021	0.157	0.264
F1 99%	Random Forest	50,000	0.0092	stratified	-0.002 (0.0008)	0.00026 (1.7e-05)	0.00025	0.02	0.169	0.272
F1 99%	Random Forest	50,000	0.0184	naive	-0.0028 (0.0007)	0.00019 (1.1e-05)	0.00018	0.018	0.221	0.302
F1 99%	Random Forest	50,000	0.0184	stratified	-0.002 (0.00072)	0.00018 (1.3e-05)	0.00018	0.018	0.215	0.311
F1 99%	Random Forest	100,000	0.0046	naive	-0.0011 (0.00058)	0.00014 (8.4e-06)	0.00014	0.017	0.114	0.184
F1 99%	Random Forest	100,000	0.0046	stratified	-0.0013 (0.00057)	0.00014 (9.5e-06)	0.00014	0.015	0.109	0.18
F1 99%	Random Forest	100,000	0.0092	naive	-0.0016 (0.00057)	0.00013 (8.2e-06)	0.00013	0.016	0.18	0.251
F1 99%	Random Forest	100,000	0.0092	stratified	0.00028 (0.00057)	0.00012 (8e-06)	0.00012	0.014	0.187	0.253
F1 99%	Random Forest	100,000	0.0184	naive	-0.0014 (0.0005)	8.7e-05 (5.6e-06)	8.5e-05	0.013	0.239	0.297
F1 99%	Random Forest	100,000	0.0184	stratified	-0.0019 (0.00053)	9.5e-05 (6.1e-06)	9.1e-05	0.013	0.231	0.286
F1 99%	Random Forest	1,000,000	0.0046	naive	-0.00061 (0.00025)	1.2e-05 (7.8e-07)	1.1e-05	0.0043	0.138	0.159
F1 99%	Random Forest	1,000,000	0.0046	stratified	-0.00026 (0.00023)	9.5e-06 (5.7e-07)	9.4e-06	0.0041	0.14	0.156
F1 99%	Random Forest	1,000,000	0.0092	naive	-0.0006 (0.00025)	9.3e-06 (5.4e-07)	8.9e-06	0.0042	0.206	0.222
F1 99%	Random Forest	1,000,000	0.0092	stratified	-0.00049 (0.00025)	9e-06 (5.8e-07)	8.8e-06	0.004	0.205	0.223

Table C.54: Bias, MSE, and Variability for F_1 score at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 90%	Ridge	5,000	0.0046	naive	0.0044 (0.00035)	0.00015 (6.9e-06)	0.00013	0.015	0.011	0.0919
F1 90%	Ridge	5,000	0.0046	stratified	4.6e-05 (0.00031)	9.9e-05 (4.6e-06)	9.9e-05	0.014	0.00351	0.0839
F1 90%	Ridge	5,000	0.0092	naive	-0.0003 (0.00041)	0.00016 (7.5e-06)	0.00016	0.018	0.05	0.137
F1 90%	Ridge	5,000	0.0092	stratified	0.0001 (0.00041)	0.00016 (7e-06)	0.00016	0.018	0.0582	0.14
F1 90%	Ridge	5,000	0.0184	naive	-0.0023 (0.00051)	0.00026 (1.2e-05)	0.00026	0.022	0.122	0.223
F1 90%	Ridge	5,000	0.0184	stratified	0.00014 (0.00051)	0.00025 (1.1e-05)	0.00025	0.021	0.127	0.23
F1 90%	Ridge	10,000	0.0046	naive	0.00011 (0.00022)	4.6e-05 (1.9e-06)	4.6e-05	0.01	0.0287	0.0737
F1 90%	Ridge	10,000	0.0046	stratified	-8.5e-05 (0.0002)	4.1e-05 (1.9e-06)	4.1e-05	0.0086	0.0285	0.0711
F1 90%	Ridge	10,000	0.0092	naive	-0.001 (0.00028)	7.4e-05 (3.2e-06)	7.3e-05	0.012	0.0713	0.125
F1 90%	Ridge	10,000	0.0092	stratified	3e-05 (0.00029)	8.1e-05 (3.5e-06)	8.1e-05	0.012	0.0719	0.126
F1 90%	Ridge	10,000	0.0184	naive	-0.0017 (0.00037)	0.00013 (5.6e-06)	0.00013	0.016	0.144	0.214
F1 90%	Ridge	10,000	0.0184	stratified	-8.4e-05 (0.00038)	0.00013 (5.9e-06)	0.00013	0.015	0.147	0.225
F1 90%	Ridge	50,000	0.0046	naive	-0.00022 (9.8e-05)	8.9e-06 (4.1e-07)	8.9e-06	0.0041	0.0437	0.0637
F1 90%	Ridge	50,000	0.0046	stratified	-0.00016 (9.2e-05)	7.8e-06 (3.6e-07)	7.8e-06	0.0037	0.0429	0.0636
F1 90%	Ridge	50,000	0.0092	naive	-0.00021 (0.00012)	1.4e-05 (6.1e-07)	1.4e-05	0.005	0.093	0.115
F1 90%	Ridge	50,000	0.0092	stratified	-3.2e-05 (0.00013)	1.4e-05 (6.3e-07)	1.4e-05	0.0053	0.0902	0.116
F1 90%	Ridge	50,000	0.0184	naive	-0.0005 (0.00017)	2.4e-05 (1.1e-06)	2.4e-05	0.0063	0.176	0.207
F1 90%	Ridge	50,000	0.0184	stratified	1e-05 (0.00017)	2.4e-05 (1.1e-06)	2.4e-05	0.0067	0.175	0.208
F1 90%	Ridge	100,000	0.0046	naive	-8.9e-05 (6.7e-05)	4e-06 (1.8e-07)	4e-06	0.0026	0.0485	0.062
F1 90%	Ridge	100,000	0.0046	stratified	8.7e-05 (7e-05)	4.4e-06 (2e-07)	4.4e-06	0.0028	0.0484	0.0609
F1 90%	Ridge	100,000	0.0092	naive	-0.00017 (9e-05)	7e-06 (3.1e-07)	6.9e-06	0.0037	0.0955	0.115
F1 90%	Ridge	100,000	0.0092	stratified	4.3e-05 (9.1e-05)	7.2e-06 (3.2e-07)	7.2e-06	0.0035	0.0972	0.113
F1 90%	Ridge	100,000	0.0184	naive	-0.00026 (0.00012)	1.1e-05 (4.7e-07)	1.1e-05	0.0046	0.183	0.205
F1 90%	Ridge	100,000	0.0184	stratified	7.5e-05 (0.00012)	1.1e-05 (5.5e-07)	1.1e-05	0.0042	0.183	0.207
F1 90%	Ridge	1,000,000	0.0046	naive	-5.7e-06 (2.8e-05)	3.4e-07 (1.4e-08)	3.4e-07	0.00076	0.0545	0.0584
F1 90%	Ridge	1,000,000	0.0046	stratified	-3.2e-05 (2.8e-05)	3.4e-07 (1.4e-08)	3.4e-07	0.00086	0.0546	0.0581
F1 90%	Ridge	1,000,000	0.0092	naive	-7.1e-05 (3.8e-05)	4.6e-07 (2.1e-08)	4.6e-07	0.00088	0.106	0.11
F1 90%	Ridge	1,000,000	0.0092	stratified	-8.6e-05 (3.7e-05)	4.5e-07 (2e-08)	4.4e-07	0.00095	0.105	0.11

Table C.55: Bias, MSE, and Variability for F_1 score at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 95%	Ridge	5,000	0.0046	naive	0.0054 (0.00066)	0.00049 (2.3e-05)	0.00046	0.029	0.0205	0.16
F1 95%	Ridge	5,000	0.0046	stratified	-0.00058 (0.0006)	0.00037 (1.6e-05)	0.00037	0.026	0	0.129
F1 95%	Ridge	5,000	0.0092	naive	-0.0024 (0.00074)	0.00055 (2.4e-05)	0.00055	0.032	0.0542	0.218
F1 95%	Ridge	5,000	0.0092	stratified	0.00019 (0.00075)	0.00057 (2.6e-05)	0.00057	0.033	0.0595	0.213
F1 95%	Ridge	5,000	0.0184	naive	-0.0049 (0.00089)	0.00081 (3.6e-05)	0.00079	0.037	0.128	0.317
F1 95%	Ridge	5,000	0.0184	stratified	-0.00012 (0.00089)	0.00076 (3.2e-05)	0.00076	0.038	0.148	0.31
F1 95%	Ridge	10,000	0.0046	naive	-0.00043 (0.00042)	0.00017 (7.4e-06)	0.00017	0.018	0.0284	0.114
F1 95%	Ridge	10,000	0.0046	stratified	-7.1e-05 (0.00039)	0.00015 (7e-06)	0.00015	0.017	0.0403	0.116
F1 95%	Ridge	10,000	0.0092	naive	-0.0023 (0.00052)	0.00026 (1.1e-05)	0.00025	0.022	0.0909	0.185
F1 95%	Ridge	10,000	0.0092	stratified	-0.00058 (0.00053)	0.00028 (1.2e-05)	0.00028	0.022	0.0909	0.19
F1 95%	Ridge	10,000	0.0184	naive	-0.0031 (0.00064)	0.00041 (1.8e-05)	0.0004	0.028	0.176	0.304
F1 95%	Ridge	10,000	0.0184	stratified	-0.00032 (0.00062)	0.00037 (1.6e-05)	0.00037	0.026	0.185	0.302
F1 95%	Ridge	50,000	0.0046	naive	-0.00052 (0.00019)	3.3e-05 (1.5e-06)	3.3e-05	0.0076	0.0574	0.099
F1 95%	Ridge	50,000	0.0046	stratified	-0.00033 (0.00018)	3e-05 (1.4e-06)	3e-05	0.0078	0.0619	0.0993
F1 95%	Ridge	50,000	0.0092	naive	-0.00069 (0.00023)	5e-05 (2.2e-06)	4.9e-05	0.0093	0.123	0.169
F1 95%	Ridge	50,000	0.0092	stratified	1.5e-05 (0.00023)	4.9e-05 (2.2e-06)	4.9e-05	0.0093	0.127	0.17
F1 95%	Ridge	50,000	0.0184	naive	-0.001 (0.00028)	7.6e-05 (3.6e-06)	7.5e-05	0.011	0.224	0.288
F1 95%	Ridge	50,000	0.0184	stratified	-0.00025 (0.00028)	7e-05 (3.1e-06)	7e-05	0.011	0.23	0.289
F1 95%	Ridge	100,000	0.0046	naive	-0.00022 (0.00013)	1.6e-05 (7.1e-07)	1.6e-05	0.0056	0.0688	0.0955
F1 95%	Ridge	100,000	0.0046	stratified	5.6e-05 (0.00013)	1.6e-05 (7.2e-07)	1.6e-05	0.0054	0.0695	0.0945
F1 95%	Ridge	100,000	0.0092	naive	-0.00047 (0.00017)	2.5e-05 (1.1e-06)	2.5e-05	0.0068	0.137	0.169
F1 95%	Ridge	100,000	0.0092	stratified	-0.00017 (0.00017)	2.5e-05 (1.1e-06)	2.5e-05	0.0063	0.137	0.167
F1 95%	Ridge	100,000	0.0184	naive	-0.00048 (0.0002)	3.3e-05 (1.4e-06)	3.3e-05	0.0077	0.239	0.275
F1 95%	Ridge	100,000	0.0184	stratified	-0.00012 (0.00021)	3.6e-05 (1.8e-06)	3.6e-05	0.0079	0.238	0.287
F1 95%	Ridge	1,000,000	0.0046	naive	-1.6e-05 (5.6e-05)	1.3e-06 (5.6e-08)	1.3e-06	0.0015	0.0824	0.0892
F1 95%	Ridge	1,000,000	0.0046	stratified	-5.3e-05 (5.5e-05)	1.4e-06 (6e-08)	1.4e-06	0.0016	0.0822	0.0894
F1 95%	Ridge	1,000,000	0.0092	naive	-0.00012 (7e-05)	1.7e-06 (7.2e-08)	1.7e-06	0.0018	0.152	0.16
F1 95%	Ridge	1,000,000	0.0092	stratified	-6.2e-05 (6.8e-05)	1.6e-06 (7.5e-08)	1.6e-06	0.0016	0.151	0.16

Table C.56: Bias, MSE, and Variability for F_1 score at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 99%	Ridge	5,000	0.0046	naive	-0.00088 (0.0019)	0.0038 (0.00016)	0.0038	0.088	0	0.325
F1 99%	Ridge	5,000	0.0046	stratified	-0.0085 (0.0018)	0.0035 (0.00014)	0.0034	0.086	0	0.314
F1 99%	Ridge	5,000	0.0092	naive	-0.016 (0.0018)	0.0036 (0.00015)	0.0033	0.077	0.0182	0.352
F1 99%	Ridge	5,000	0.0092	stratified	-0.0078 (0.0018)	0.0035 (0.00015)	0.0035	0.078	0.02	0.39
F1 99%	Ridge	5,000	0.0184	naive	-0.0087 (0.0016)	0.0024 (0.00011)	0.0024	0.066	0.0858	0.4
F1 99%	Ridge	5,000	0.0184	stratified	-0.009 (0.0016)	0.0025 (0.00011)	0.0024	0.067	0.0971	0.441
F1 99%	Ridge	10,000	0.0046	naive	-0.006 (0.0012)	0.0015 (6.3e-05)	0.0015	0.052	0.0276	0.242
F1 99%	Ridge	10,000	0.0046	stratified	-0.0027 (0.0013)	0.0018 (7.8e-05)	0.0018	0.056	0.0235	0.278
F1 99%	Ridge	10,000	0.0092	naive	-0.0093 (0.0013)	0.0018 (7.3e-05)	0.0017	0.059	0.0808	0.331
F1 99%	Ridge	10,000	0.0092	stratified	-0.0047 (0.0013)	0.0018 (7.9e-05)	0.0018	0.057	0.0881	0.336
F1 99%	Ridge	10,000	0.0184	naive	-0.0057 (0.0012)	0.0013 (5.9e-05)	0.0013	0.046	0.131	0.371
F1 99%	Ridge	10,000	0.0184	stratified	-0.0065 (0.0011)	0.0012 (5e-05)	0.0012	0.048	0.158	0.354
F1 99%	Ridge	50,000	0.0046	naive	-0.0027 (0.00054)	0.00029 (1.3e-05)	0.00028	0.022	0.0918	0.209
F1 99%	Ridge	50,000	0.0046	stratified	-0.0023 (0.00056)	0.00031 (1.3e-05)	0.00031	0.024	0.094	0.199
F1 99%	Ridge	50,000	0.0092	naive	-0.0022 (0.00056)	0.0003 (1.4e-05)	0.0003	0.023	0.16	0.269
F1 99%	Ridge	50,000	0.0092	stratified	-0.0011 (0.00059)	0.00033 (1.5e-05)	0.00033	0.024	0.165	0.27
F1 99%	Ridge	50,000	0.0184	naive	-0.0023 (0.00049)	0.00023 (9.8e-06)	0.00022	0.02	0.226	0.318
F1 99%	Ridge	50,000	0.0184	stratified	-0.0023 (0.00049)	0.00022 (1e-05)	0.00021	0.019	0.216	0.319
F1 99%	Ridge	100,000	0.0046	naive	-0.0012 (0.0004)	0.00015 (6.4e-06)	0.00015	0.016	0.115	0.187
F1 99%	Ridge	100,000	0.0046	stratified	-0.00015 (0.0004)	0.00015 (6.4e-06)	0.00015	0.017	0.121	0.189
F1 99%	Ridge	100,000	0.0092	naive	-0.0021 (0.00042)	0.00016 (6.9e-06)	0.00016	0.017	0.187	0.262
F1 99%	Ridge	100,000	0.0092	stratified	-0.0011 (0.00041)	0.00016 (6.9e-06)	0.00016	0.017	0.185	0.264
F1 99%	Ridge	100,000	0.0184	naive	-0.0012 (0.00037)	0.00011 (4.9e-06)	0.00011	0.014	0.243	0.31
F1 99%	Ridge	100,000	0.0184	stratified	-0.001 (0.00037)	0.00011 (5e-06)	0.00011	0.014	0.237	0.306
F1 99%	Ridge	1,000,000	0.0046	naive	-0.00018 (0.00018)	1.4e-05 (6.3e-07)	1.4e-05	0.0049	0.149	0.172
F1 99%	Ridge	1,000,000	0.0046	stratified	-8.1e-05 (0.00017)	1.3e-05 (5.9e-07)	1.3e-05	0.005	0.148	0.173
F1 99%	Ridge	1,000,000	0.0092	naive	-0.0003 (0.00017)	1e-05 (4.6e-07)	1e-05	0.0041	0.223	0.244
F1 99%	Ridge	1,000,000	0.0092	stratified	-0.00033 (0.00017)	1.1e-05 (4.7e-07)	1e-05	0.0043	0.223	0.243

Table C.57: Bias, MSE, and Variability for F_1 score at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 90%	Logistic	5,000	0.0046	naive	-0.00075 (0.00015)	1.5e-05 (1.1e-06)	1.4e-05	0.0027	0	0.0325
F1 90%	Logistic	5,000	0.0046	stratified	-0.0013 (0.00014)	1.5e-05 (8.6e-07)	1.3e-05	0.0021	0	0.0333
F1 90%	Logistic	5,000	0.0092	naive	-0.0058 (0.00041)	0.00019 (7.4e-06)	0.00015	0.017	0.00725	0.0801
F1 90%	Logistic	5,000	0.0092	stratified	-0.0036 (0.00042)	0.00018 (7.4e-06)	0.00017	0.018	0.00775	0.0804
F1 90%	Logistic	5,000	0.0184	naive	-0.0075 (0.00063)	0.00044 (2.7e-05)	0.00038	0.025	0.0343	0.176
F1 90%	Logistic	5,000	0.0184	stratified	-0.0063 (0.00065)	0.00048 (3e-05)	0.00044	0.024	0.0345	0.185
F1 90%	Logistic	10,000	0.0046	naive	-0.0021 (0.00023)	4.8e-05 (2e-06)	4.4e-05	0.0091	0.00476	0.0466
F1 90%	Logistic	10,000	0.0046	stratified	-0.0016 (0.00022)	4.6e-05 (1.9e-06)	4.3e-05	0.009	0.00698	0.0428
F1 90%	Logistic	10,000	0.0092	naive	-0.004 (0.00034)	0.00012 (5.1e-06)	0.0001	0.013	0.0364	0.0999
F1 90%	Logistic	10,000	0.0092	stratified	-0.0035 (0.00033)	0.0001 (4.6e-06)	9.1e-05	0.013	0.0341	0.0984
F1 90%	Logistic	10,000	0.0184	naive	-0.0047 (0.0004)	0.00017 (7.4e-06)	0.00015	0.017	0.121	0.19
F1 90%	Logistic	10,000	0.0184	stratified	-0.0045 (0.00041)	0.00017 (8.2e-06)	0.00015	0.017	0.106	0.193
F1 90%	Logistic	50,000	0.0046	naive	-0.00088 (0.0001)	9.9e-06 (4.4e-07)	9.1e-06	0.0038	0.0385	0.0572
F1 90%	Logistic	50,000	0.0046	stratified	-0.00076 (0.0001)	1e-05 (4.5e-07)	9.4e-06	0.004	0.0383	0.0587
F1 90%	Logistic	50,000	0.0092	naive	-0.00084 (0.00014)	1.8e-05 (8e-07)	1.7e-05	0.0056	0.0881	0.113
F1 90%	Logistic	50,000	0.0092	stratified	-0.00038 (0.00013)	1.5e-05 (6.8e-07)	1.5e-05	0.0054	0.0879	0.114
F1 90%	Logistic	50,000	0.0184	naive	-0.0009 (0.00017)	2.7e-05 (1.2e-06)	2.6e-05	0.007	0.17	0.205
F1 90%	Logistic	50,000	0.0184	stratified	-0.00056 (0.00017)	2.6e-05 (1.1e-06)	2.6e-05	0.0071	0.173	0.206
F1 90%	Logistic	100,000	0.0046	naive	-0.00031 (7.2e-05)	4.9e-06 (2.3e-07)	4.8e-06	0.0028	0.0451	0.0605
F1 90%	Logistic	100,000	0.0046	stratified	-0.0003 (7.1e-05)	4.5e-06 (2e-07)	4.5e-06	0.0029	0.0461	0.0589
F1 90%	Logistic	100,000	0.0092	naive	-0.0004 (9.2e-05)	7.5e-06 (3.4e-07)	7.4e-06	0.0036	0.0951	0.111
F1 90%	Logistic	100,000	0.0092	stratified	-0.00029 (9.5e-05)	7.8e-06 (3.6e-07)	7.8e-06	0.0039	0.0922	0.112
F1 90%	Logistic	100,000	0.0184	naive	-0.00032 (0.00012)	1.2e-05 (5e-07)	1.2e-05	0.0047	0.182	0.202
F1 90%	Logistic	100,000	0.0184	stratified	-0.00034 (0.00012)	1.1e-05 (4.8e-07)	1.1e-05	0.0047	0.183	0.202
F1 90%	Logistic	1,000,000	0.0046	naive	-4e-05 (2.8e-05)	3.3e-07 (1.5e-08)	3.3e-07	0.0008	0.0545	0.0582
F1 90%	Logistic	1,000,000	0.0046	stratified	-1.1e-05 (2.9e-05)	3.2e-07 (1.3e-08)	3.2e-07	0.00078	0.0547	0.058
F1 90%	Logistic	1,000,000	0.0092	naive	-1.1e-05 (3.8e-05)	4.7e-07 (2.2e-08)	4.7e-07	0.00093	0.105	0.11
F1 90%	Logistic	1,000,000	0.0092	stratified	3.9e-06 (3.8e-05)	4.7e-07 (2.1e-08)	4.7e-07	0.00096	0.105	0.11

Table C.58: Bias, MSE, and Variability for F_1 score at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 95%	Logistic	5,000	0.0046	naive	-0.0021 (0.00025)	4.6e-05 (3.1e-06)	4.2e-05	0.0049	0	0.0501
F1 95%	Logistic	5,000	0.0046	stratified	-0.0026 (0.00023)	4.4e-05 (2.2e-06)	3.8e-05	0.0045	0	0.042
F1 95%	Logistic	5,000	0.0092	naive	-0.0089 (0.00065)	0.00049 (1.9e-05)	0.00042	0.028	0.00606	0.126
F1 95%	Logistic	5,000	0.0092	stratified	-0.0054 (0.00067)	0.00047 (1.9e-05)	0.00044	0.029	0	0.119
F1 95%	Logistic	5,000	0.0184	naive	-0.011 (0.00095)	0.001 (5.5e-05)	0.0009	0.04	0.0343	0.245
F1 95%	Logistic	5,000	0.0184	stratified	-0.009 (0.00098)	0.0011 (6.2e-05)	0.001	0.041	0.0342	0.25
F1 95%	Logistic	10,000	0.0046	naive	-0.004 (0.00037)	0.00014 (5.8e-06)	0.00012	0.015	0	0.0688
F1 95%	Logistic	10,000	0.0046	stratified	-0.0024 (0.00037)	0.00013 (5.3e-06)	0.00013	0.015	0.00618	0.0666
F1 95%	Logistic	10,000	0.0092	naive	-0.0069 (0.00056)	0.00034 (1.4e-05)	0.00029	0.023	0.0333	0.157
F1 95%	Logistic	10,000	0.0092	stratified	-0.0054 (0.00056)	0.00031 (1.4e-05)	0.00028	0.022	0.0463	0.164
F1 95%	Logistic	10,000	0.0184	naive	-0.0078 (0.00065)	0.00047 (2e-05)	0.00041	0.028	0.141	0.263
F1 95%	Logistic	10,000	0.0184	stratified	-0.0067 (0.00065)	0.00045 (2e-05)	0.00041	0.027	0.138	0.277
F1 95%	Logistic	50,000	0.0046	naive	-0.0017 (0.00019)	3.6e-05 (1.6e-06)	3.3e-05	0.0075	0.0527	0.0877
F1 95%	Logistic	50,000	0.0046	stratified	-0.0015 (0.00019)	3.7e-05 (1.6e-06)	3.5e-05	0.0077	0.0522	0.091
F1 95%	Logistic	50,000	0.0092	naive	-0.0016 (0.00024)	5.9e-05 (2.9e-06)	5.6e-05	0.0097	0.118	0.169
F1 95%	Logistic	50,000	0.0092	stratified	-0.00081 (0.00023)	5.3e-05 (2.4e-06)	5.2e-05	0.0097	0.12	0.168
F1 95%	Logistic	50,000	0.0184	naive	-0.0018 (0.00028)	7.7e-05 (3.3e-06)	7.4e-05	0.012	0.226	0.28
F1 95%	Logistic	50,000	0.0184	stratified	-0.00089 (0.00029)	8e-05 (3.7e-06)	7.9e-05	0.012	0.223	0.287
F1 95%	Logistic	100,000	0.0046	naive	-0.00069 (0.00014)	1.8e-05 (8.1e-07)	1.8e-05	0.0056	0.0671	0.094
F1 95%	Logistic	100,000	0.0046	stratified	-0.00056 (0.00014)	1.7e-05 (7.7e-07)	1.7e-05	0.0056	0.0665	0.0947
F1 95%	Logistic	100,000	0.0092	naive	-0.00082 (0.00018)	2.8e-05 (1.2e-06)	2.8e-05	0.0072	0.134	0.166
F1 95%	Logistic	100,000	0.0092	stratified	-0.00061 (0.00018)	2.7e-05 (1.2e-06)	2.7e-05	0.0066	0.135	0.166
F1 95%	Logistic	100,000	0.0184	naive	-0.00048 (0.0002)	3.7e-05 (1.5e-06)	3.6e-05	0.0084	0.242	0.277
F1 95%	Logistic	100,000	0.0184	stratified	-0.00051 (0.0002)	3.3e-05 (1.4e-06)	3.3e-05	0.0079	0.239	0.272
F1 95%	Logistic	1,000,000	0.0046	naive	-5e-05 (5.6e-05)	1.3e-06 (5.9e-08)	1.3e-06	0.0016	0.082	0.0899
F1 95%	Logistic	1,000,000	0.0046	stratified	-8.8e-06 (5.6e-05)	1.4e-06 (5.9e-08)	1.4e-06	0.0015	0.0825	0.0891
F1 95%	Logistic	1,000,000	0.0092	naive	-2.9e-06 (7e-05)	1.8e-06 (8.5e-08)	1.8e-06	0.0017	0.152	0.162
F1 95%	Logistic	1,000,000	0.0092	stratified	1.9e-05 (7.2e-05)	1.8e-06 (8e-08)	1.8e-06	0.0019	0.152	0.161

Table C.59: Bias, MSE, and Variability for F_1 score at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F1 99%	Logistic	5,000	0.0046	naive	0.0012 (0.00074)	0.00053 (3.3e-05)	0.00053	0.03	0	0.145
F1 99%	Logistic	5,000	0.0046	stratified	-0.00027 (0.00071)	0.0005 (2.7e-05)	0.0005	0.033	0	0.127
F1 99%	Logistic	5,000	0.0092	naive	-0.014 (0.0011)	0.0015 (6.2e-05)	0.0013	0.048	0	0.21
F1 99%	Logistic	5,000	0.0092	stratified	-0.0093 (0.0012)	0.0014 (5.4e-05)	0.0013	0.051	0	0.22
F1 99%	Logistic	5,000	0.0184	naive	-0.014 (0.0014)	0.002 (8.4e-05)	0.0018	0.06	0.0408	0.302
F1 99%	Logistic	5,000	0.0184	stratified	-0.013 (0.0014)	0.0021 (9.3e-05)	0.0019	0.058	0.0246	0.299
F1 99%	Logistic	10,000	0.0046	naive	-0.0099 (0.00084)	0.00075 (2.8e-05)	0.00066	0.034	0	0.141
F1 99%	Logistic	10,000	0.0046	stratified	-0.0055 (0.00089)	0.00079 (3.4e-05)	0.00076	0.037	0	0.175
F1 99%	Logistic	10,000	0.0092	naive	-0.013 (0.001)	0.0013 (5.3e-05)	0.0011	0.045	0.0306	0.251
F1 99%	Logistic	10,000	0.0092	stratified	-0.01 (0.0011)	0.0014 (5.8e-05)	0.0013	0.048	0.0105	0.252
F1 99%	Logistic	10,000	0.0184	naive	-0.011 (0.0011)	0.0012 (5.3e-05)	0.0011	0.043	0.104	0.344
F1 99%	Logistic	10,000	0.0184	stratified	-0.012 (0.0011)	0.0012 (5.2e-05)	0.0011	0.044	0.104	0.306
F1 99%	Logistic	50,000	0.0046	naive	-0.0044 (0.00054)	0.00031 (1.3e-05)	0.00029	0.024	0.0728	0.182
F1 99%	Logistic	50,000	0.0046	stratified	-0.004 (0.00054)	0.00032 (1.4e-05)	0.0003	0.024	0.0743	0.191
F1 99%	Logistic	50,000	0.0092	naive	-0.0046 (0.00057)	0.00035 (1.5e-05)	0.00033	0.024	0.152	0.26
F1 99%	Logistic	50,000	0.0092	stratified	-0.003 (0.00057)	0.00032 (1.4e-05)	0.00031	0.024	0.15	0.269
F1 99%	Logistic	50,000	0.0184	naive	-0.0031 (0.00051)	0.00024 (1e-05)	0.00023	0.021	0.212	0.316
F1 99%	Logistic	50,000	0.0184	stratified	-0.0025 (0.00052)	0.00024 (1.1e-05)	0.00024	0.021	0.213	0.318
F1 99%	Logistic	100,000	0.0046	naive	-0.0023 (0.00041)	0.00016 (7.3e-06)	0.00016	0.017	0.0987	0.188
F1 99%	Logistic	100,000	0.0046	stratified	-0.0015 (0.00041)	0.00017 (7.3e-06)	0.00016	0.017	0.107	0.189
F1 99%	Logistic	100,000	0.0092	naive	-0.0024 (0.00043)	0.00017 (7.6e-06)	0.00017	0.017	0.182	0.262
F1 99%	Logistic	100,000	0.0092	stratified	-0.0025 (0.00042)	0.00017 (7.8e-06)	0.00016	0.017	0.175	0.263
F1 99%	Logistic	100,000	0.0184	naive	-0.0008 (0.00036)	0.00011 (5.1e-06)	0.00011	0.014	0.233	0.313
F1 99%	Logistic	100,000	0.0184	stratified	-0.0015 (0.00036)	0.00011 (4.8e-06)	0.0001	0.013	0.241	0.304
F1 99%	Logistic	1,000,000	0.0046	naive	-0.00021 (0.00018)	1.4e-05 (6.1e-07)	1.4e-05	0.0049	0.15	0.172
F1 99%	Logistic	1,000,000	0.0046	stratified	-0.00037 (0.00018)	1.4e-05 (6.1e-07)	1.3e-05	0.0049	0.151	0.174
F1 99%	Logistic	1,000,000	0.0092	naive	4e-06 (0.00017)	1.2e-05 (5.4e-07)	1.2e-05	0.0048	0.224	0.246
F1 99%	Logistic	1,000,000	0.0092	stratified	0.0001 (0.00018)	1.2e-05 (5.1e-07)	1.2e-05	0.0045	0.222	0.243

Table C.60: Bias, MSE, and Variability for F_1 score at the 99th percentile, logistic regression.

C.9 $F_{0.5}$ *Score*

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 90%	Random Forest	5,000	0.0046	naive	0.011 (0.00031)	0.00017 (8.9e-06)	4.9e-05	0.0085	0.023	0.0703
F0.5 90%	Random Forest	5,000	0.0046	stratified	0.0059 (0.00023)	6.3e-05 (3.6e-06)	2.8e-05	0.0067	0.0222	0.0542
F0.5 90%	Random Forest	5,000	0.0092	naive	0.0042 (0.0004)	9.9e-05 (7.1e-06)	8.1e-05	0.012	0.0424	0.0967
F0.5 90%	Random Forest	5,000	0.0092	stratified	0.0014 (0.00035)	6.5e-05 (4e-06)	6.3e-05	0.011	0.0397	0.0874
F0.5 90%	Random Forest	5,000	0.0184	naive	-0.00037 (0.00051)	0.00012 (7e-06)	0.00012	0.014	0.0949	0.156
F0.5 90%	Random Forest	5,000	0.0184	stratified	0.001 (0.00051)	0.00012 (7.3e-06)	0.00012	0.015	0.0898	0.158
F0.5 90%	Random Forest	10,000	0.0046	naive	0.0025 (0.00021)	2.7e-05 (1.7e-06)	2e-05	0.0061	0.023	0.0485
F0.5 90%	Random Forest	10,000	0.0046	stratified	0.00094 (0.00018)	1.6e-05 (9.9e-07)	1.5e-05	0.0053	0.022	0.045
F0.5 90%	Random Forest	10,000	0.0092	naive	-9e-06 (0.00026)	3.5e-05 (2.3e-06)	3.5e-05	0.0077	0.0463	0.0801
F0.5 90%	Random Forest	10,000	0.0092	stratified	0.00048 (0.00025)	3.1e-05 (1.9e-06)	3.1e-05	0.0074	0.0489	0.0809
F0.5 90%	Random Forest	10,000	0.0184	naive	-0.00094 (0.00038)	6.7e-05 (4e-06)	6.6e-05	0.011	0.104	0.152
F0.5 90%	Random Forest	10,000	0.0184	stratified	0.0012 (0.00036)	5.9e-05 (3.8e-06)	5.8e-05	0.0099	0.108	0.151
F0.5 90%	Random Forest	50,000	0.0046	naive	-3.8e-05 (8.4e-05)	3.1e-06 (1.9e-07)	3.1e-06	0.0025	0.0288	0.0389
F0.5 90%	Random Forest	50,000	0.0046	stratified	-2.9e-05 (8.3e-05)	3e-06 (1.9e-07)	3e-06	0.0024	0.0283	0.0395
F0.5 90%	Random Forest	50,000	0.0092	naive	-0.0002 (0.00012)	5.9e-06 (3.7e-07)	5.8e-06	0.0034	0.0608	0.0761
F0.5 90%	Random Forest	50,000	0.0092	stratified	-6.6e-05 (0.00012)	6.1e-06 (3.9e-07)	6.2e-06	0.0033	0.0614	0.0764
F0.5 90%	Random Forest	50,000	0.0184	naive	-0.00057 (0.00017)	1.2e-05 (7.4e-07)	1.1e-05	0.0044	0.121	0.142
F0.5 90%	Random Forest	50,000	0.0184	stratified	-0.00021 (0.00017)	1.2e-05 (7.1e-07)	1.2e-05	0.0044	0.123	0.143
F0.5 90%	Random Forest	100,000	0.0046	naive	-6.6e-05 (6.2e-05)	1.6e-06 (9.9e-08)	1.6e-06	0.0016	0.0312	0.0392
F0.5 90%	Random Forest	100,000	0.0046	stratified	-2.3e-05 (6.2e-05)	1.7e-06 (1e-07)	1.7e-06	0.0017	0.0318	0.0392
F0.5 90%	Random Forest	100,000	0.0092	naive	-5.7e-05 (8.2e-05)	2.7e-06 (1.6e-07)	2.7e-06	0.0022	0.0643	0.0738
F0.5 90%	Random Forest	100,000	0.0092	stratified	1.3e-05 (8.7e-05)	3.1e-06 (2e-07)	3.1e-06	0.0024	0.0648	0.0763
F0.5 90%	Random Forest	100,000	0.0184	naive	-0.00028 (0.00012)	5.6e-06 (3.5e-07)	5.5e-06	0.0031	0.127	0.141
F0.5 90%	Random Forest	100,000	0.0184	stratified	-0.00031 (0.00012)	5.5e-06 (3.4e-07)	5.4e-06	0.0032	0.128	0.142
F0.5 90%	Random Forest	1,000,000	0.0046	naive	-7.6e-05 (2.6e-05)	1.4e-07 (8.4e-09)	1.3e-07	0.00049	0.0347	0.0369
F0.5 90%	Random Forest	1,000,000	0.0046	stratified	-7.8e-05 (2.6e-05)	1.3e-07 (8.4e-09)	1.3e-07	0.00048	0.0345	0.037
F0.5 90%	Random Forest	1,000,000	0.0092	naive	-0.00016 (3.6e-05)	2.3e-07 (1.4e-08)	2e-07	0.00061	0.0688	0.0715
F0.5 90%	Random Forest	1,000,000	0.0092	stratified	-0.0002 (3.8e-05)	2.2e-07 (1.3e-08)	1.9e-07	0.00056	0.0689	0.0713

Table C.61: Bias, MSE, and Variability for $F_{0.5}$ score at the 90th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 95%	Random Forest	5,000	0.0046	naive	0.025 (0.00063)	0.00084 (4.1e-05)	0.0002	0.018	0.0409	0.134
F0.5 95%	Random Forest	5,000	0.0046	stratified	0.018 (0.00044)	0.00043 (2e-05)	0.0001	0.013	0.0399	0.101
F0.5 95%	Random Forest	5,000	0.0092	naive	0.013 (0.00078)	0.00045 (3.1e-05)	0.00029	0.021	0.0605	0.174
F0.5 95%	Random Forest	5,000	0.0092	stratified	0.008 (0.00068)	0.00029 (1.9e-05)	0.00023	0.019	0.0561	0.147
F0.5 95%	Random Forest	5,000	0.0184	naive	0.00072 (0.00097)	0.0004 (2.4e-05)	0.0004	0.027	0.124	0.242
F0.5 95%	Random Forest	5,000	0.0184	stratified	0.0024 (0.00097)	0.00043 (2.6e-05)	0.00043	0.027	0.123	0.243
F0.5 95%	Random Forest	10,000	0.0046	naive	0.0076 (0.00042)	0.00014 (9.2e-06)	7.9e-05	0.012	0.0312	0.0964
F0.5 95%	Random Forest	10,000	0.0046	stratified	0.0042 (0.00036)	7.5e-05 (4.7e-06)	5.8e-05	0.011	0.0319	0.0758
F0.5 95%	Random Forest	10,000	0.0092	naive	0.00035 (0.00052)	0.00013 (8.7e-06)	0.00013	0.015	0.0636	0.139
F0.5 95%	Random Forest	10,000	0.0092	stratified	0.0014 (0.0005)	0.00012 (7.9e-06)	0.00012	0.014	0.0644	0.132
F0.5 95%	Random Forest	10,000	0.0184	naive	-0.0019 (0.00064)	0.0002 (1.2e-05)	0.00019	0.017	0.148	0.231
F0.5 95%	Random Forest	10,000	0.0184	stratified	0.00064 (0.00066)	0.0002 (1.3e-05)	0.0002	0.019	0.155	0.237
F0.5 95%	Random Forest	50,000	0.0046	naive	-0.00012 (0.00017)	1.2e-05 (7.2e-07)	1.2e-05	0.0044	0.0428	0.0634
F0.5 95%	Random Forest	50,000	0.0046	stratified	8.7e-05 (0.00017)	1.2e-05 (7.8e-07)	1.2e-05	0.0046	0.0426	0.0635
F0.5 95%	Random Forest	50,000	0.0092	naive	-0.00046 (0.00024)	2.4e-05 (1.5e-06)	2.4e-05	0.0066	0.0906	0.117
F0.5 95%	Random Forest	50,000	0.0092	stratified	-0.00027 (0.00025)	2.5e-05 (1.6e-06)	2.5e-05	0.0068	0.0885	0.12
F0.5 95%	Random Forest	50,000	0.0184	naive	-0.0014 (0.00032)	4.2e-05 (2.6e-06)	4e-05	0.008	0.177	0.214
F0.5 95%	Random Forest	50,000	0.0184	stratified	-0.00072 (0.00033)	4.5e-05 (2.8e-06)	4.5e-05	0.0089	0.175	0.214
F0.5 95%	Random Forest	100,000	0.0046	naive	-0.00019 (0.00013)	6.5e-06 (4.5e-07)	6.5e-06	0.003	0.0464	0.0619
F0.5 95%	Random Forest	100,000	0.0046	stratified	-0.00022 (0.00013)	7.1e-06 (4.3e-07)	7.1e-06	0.0036	0.0473	0.0633
F0.5 95%	Random Forest	100,000	0.0092	naive	-0.00015 (0.00016)	1e-05 (6.1e-07)	1e-05	0.0046	0.097	0.115
F0.5 95%	Random Forest	100,000	0.0092	stratified	-1.6e-05 (0.00017)	1.1e-05 (6.8e-07)	1.1e-05	0.0046	0.0945	0.116
F0.5 95%	Random Forest	100,000	0.0184	naive	-0.00065 (0.00023)	2.2e-05 (1.5e-06)	2.2e-05	0.0064	0.182	0.215
F0.5 95%	Random Forest	100,000	0.0184	stratified	-0.00028 (0.00024)	2.2e-05 (1.3e-06)	2.2e-05	0.0063	0.186	0.212
F0.5 95%	Random Forest	1,000,000	0.0046	naive	-0.00016 (5.2e-05)	5.9e-07 (3.6e-08)	5.7e-07	0.001	0.0531	0.0576
F0.5 95%	Random Forest	1,000,000	0.0046	stratified	-0.0002 (5e-05)	5e-07 (2.9e-08)	4.6e-07	0.00091	0.0536	0.0576
F0.5 95%	Random Forest	1,000,000	0.0092	naive	-0.00023 (7.5e-05)	8.9e-07 (5.7e-08)	8.4e-07	0.0013	0.104	0.109
F0.5 95%	Random Forest	1,000,000	0.0092	stratified	-0.0003 (7.2e-05)	8.7e-07 (5.5e-08)	7.8e-07	0.0012	0.104	0.109

Table C.62: Bias, MSE, and Variability for $F_{0.5}$ score at the 95th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 99%	Random Forest	5,000	0.0046	naive	0.16 (0.0032)	0.03 (0.0012)	0.0051	0.086	0.102	0.52
F0.5 99%	Random Forest	5,000	0.0046	stratified	0.16 (0.003)	0.031 (0.0012)	0.0046	0.08	0.098	0.551
F0.5 99%	Random Forest	5,000	0.0092	naive	0.11 (0.0027)	0.016 (0.00071)	0.0036	0.075	0.111	0.56
F0.5 99%	Random Forest	5,000	0.0092	stratified	0.11 (0.0027)	0.016 (0.00067)	0.0035	0.084	0.146	0.557
F0.5 99%	Random Forest	5,000	0.0184	naive	0.044 (0.0025)	0.0051 (0.00034)	0.0032	0.077	0.214	0.596
F0.5 99%	Random Forest	5,000	0.0184	stratified	0.045 (0.0022)	0.0044 (0.00024)	0.0024	0.066	0.197	0.513
F0.5 99%	Random Forest	10,000	0.0046	naive	0.064 (0.0015)	0.0052 (0.00022)	0.0011	0.045	0.0875	0.27
F0.5 99%	Random Forest	10,000	0.0046	stratified	0.055 (0.0014)	0.004 (0.00017)	0.00093	0.039	0.0735	0.255
F0.5 99%	Random Forest	10,000	0.0092	naive	0.028 (0.0017)	0.0022 (0.00014)	0.0015	0.054	0.111	0.362
F0.5 99%	Random Forest	10,000	0.0092	stratified	0.027 (0.0017)	0.0022 (0.00013)	0.0014	0.054	0.121	0.359
F0.5 99%	Random Forest	10,000	0.0184	naive	-0.003 (0.0018)	0.0016 (0.00011)	0.0016	0.052	0.197	0.46
F0.5 99%	Random Forest	10,000	0.0184	stratified	0.0011 (0.0018)	0.0017 (9.9e-05)	0.0017	0.06	0.185	0.432
F0.5 99%	Random Forest	50,000	0.0046	naive	-0.00042 (0.00067)	0.0002 (1.3e-05)	0.0002	0.017	0.073	0.157
F0.5 99%	Random Forest	50,000	0.0046	stratified	-4.8e-05 (0.00066)	0.00019 (1.2e-05)	0.00019	0.02	0.0859	0.165
F0.5 99%	Random Forest	50,000	0.0092	naive	-0.0013 (0.00076)	0.00025 (1.6e-05)	0.00025	0.021	0.155	0.258
F0.5 99%	Random Forest	50,000	0.0092	stratified	-0.0011 (0.00077)	0.00025 (1.6e-05)	0.00025	0.02	0.166	0.27
F0.5 99%	Random Forest	50,000	0.0184	naive	-0.003 (0.0008)	0.00027 (1.7e-05)	0.00026	0.021	0.269	0.368
F0.5 99%	Random Forest	50,000	0.0184	stratified	-0.0018 (0.00082)	0.00027 (1.8e-05)	0.00027	0.021	0.263	0.38
F0.5 99%	Random Forest	100,000	0.0046	naive	-0.00057 (0.00048)	9.6e-05 (5.7e-06)	9.6e-05	0.014	0.0935	0.152
F0.5 99%	Random Forest	100,000	0.0046	stratified	-0.00088 (0.00047)	9.3e-05 (6.4e-06)	9.3e-05	0.012	0.0893	0.148
F0.5 99%	Random Forest	100,000	0.0092	naive	-0.0012 (0.00055)	0.00013 (7.9e-06)	0.00012	0.016	0.176	0.246
F0.5 99%	Random Forest	100,000	0.0092	stratified	0.0005 (0.00056)	0.00012 (7.9e-06)	0.00012	0.014	0.183	0.248
F0.5 99%	Random Forest	100,000	0.0184	naive	-0.0017 (0.00057)	0.00013 (8.2e-06)	0.00012	0.016	0.292	0.362
F0.5 99%	Random Forest	100,000	0.0184	stratified	-0.0018 (0.00059)	0.00014 (8.8e-06)	0.00013	0.016	0.281	0.349
F0.5 99%	Random Forest	1,000,000	0.0046	naive	-0.00045 (0.00021)	7.7e-06 (5.2e-07)	7.5e-06	0.0035	0.113	0.13
F0.5 99%	Random Forest	1,000,000	0.0046	stratified	-0.00018 (0.00019)	6.3e-06 (3.8e-07)	6.3e-06	0.0033	0.114	0.128
F0.5 99%	Random Forest	1,000,000	0.0092	naive	-0.00054 (0.00024)	8.8e-06 (5.1e-07)	8.5e-06	0.0041	0.201	0.216
F0.5 99%	Random Forest	1,000,000	0.0092	stratified	-0.00045 (0.00024)	8.6e-06 (5.5e-07)	8.4e-06	0.0039	0.2	0.218

Table C.63: Bias, MSE, and Variability for $F_{0.5}$ score at the 99th percentile, random forest.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 90%	Ridge	5,000	0.0046	naive	0.012 (0.00023)	0.00019 (6.5e-06)	5.4e-05	0.01	0.0236	0.0707
F0.5 90%	Ridge	5,000	0.0046	stratified	0.0062 (0.00017)	6.6e-05 (2.5e-06)	2.7e-05	0.007	0.0225	0.0617
F0.5 90%	Ridge	5,000	0.0092	naive	0.0046 (0.00029)	0.00011 (5.8e-06)	8.4e-05	0.012	0.0406	0.111
F0.5 90%	Ridge	5,000	0.0092	stratified	0.0018 (0.00026)	6.8e-05 (3e-06)	6.5e-05	0.011	0.0389	0.0931
F0.5 90%	Ridge	5,000	0.0184	naive	-1.4e-05 (0.00036)	0.00013 (5.8e-06)	0.00013	0.015	0.0882	0.166
F0.5 90%	Ridge	5,000	0.0184	stratified	0.00051 (0.00037)	0.00013 (5.5e-06)	0.00013	0.015	0.09	0.163
F0.5 90%	Ridge	10,000	0.0046	naive	0.0025 (0.00015)	2.7e-05 (1.4e-06)	2.1e-05	0.0059	0.0208	0.0526
F0.5 90%	Ridge	10,000	0.0046	stratified	0.00073 (0.00013)	1.6e-05 (7.4e-07)	1.6e-05	0.0053	0.0205	0.046
F0.5 90%	Ridge	10,000	0.0092	naive	-1.8e-05 (0.00019)	3.3e-05 (1.4e-06)	3.3e-05	0.0079	0.0479	0.0836
F0.5 90%	Ridge	10,000	0.0092	stratified	9.6e-05 (0.00019)	3.6e-05 (1.5e-06)	3.6e-05	0.0079	0.0486	0.0838
F0.5 90%	Ridge	10,000	0.0184	naive	-0.00058 (0.00026)	6.5e-05 (2.8e-06)	6.4e-05	0.011	0.102	0.152
F0.5 90%	Ridge	10,000	0.0184	stratified	0.00013 (0.00027)	6.6e-05 (3e-06)	6.6e-05	0.011	0.104	0.16
F0.5 90%	Ridge	50,000	0.0046	naive	-0.0001 (6.3e-05)	3.7e-06 (1.7e-07)	3.7e-06	0.0027	0.0283	0.0412
F0.5 90%	Ridge	50,000	0.0046	stratified	-9.8e-05 (5.9e-05)	3.2e-06 (1.5e-07)	3.2e-06	0.0024	0.0277	0.0411
F0.5 90%	Ridge	50,000	0.0092	naive	-6e-05 (8.3e-05)	6.3e-06 (2.7e-07)	6.3e-06	0.0034	0.0621	0.077
F0.5 90%	Ridge	50,000	0.0092	stratified	-5.4e-07 (8.4e-05)	6.3e-06 (2.8e-07)	6.3e-06	0.0035	0.0602	0.0772
F0.5 90%	Ridge	50,000	0.0184	naive	-0.00023 (0.00012)	1.2e-05 (5.7e-07)	1.2e-05	0.0044	0.124	0.147
F0.5 90%	Ridge	50,000	0.0184	stratified	5.5e-05 (0.00012)	1.2e-05 (5.4e-07)	1.2e-05	0.0048	0.124	0.147
F0.5 90%	Ridge	100,000	0.0046	naive	-3.4e-05 (4.4e-05)	1.7e-06 (7.6e-08)	1.7e-06	0.0017	0.0314	0.0401
F0.5 90%	Ridge	100,000	0.0046	stratified	6e-05 (4.5e-05)	1.8e-06 (8.4e-08)	1.8e-06	0.0018	0.0313	0.0394
F0.5 90%	Ridge	100,000	0.0092	naive	-7.1e-05 (6e-05)	3.1e-06 (1.4e-07)	3.1e-06	0.0024	0.0637	0.0766
F0.5 90%	Ridge	100,000	0.0092	stratified	3.9e-05 (6.1e-05)	3.2e-06 (1.4e-07)	3.2e-06	0.0023	0.0648	0.0756
F0.5 90%	Ridge	100,000	0.0184	naive	-0.00011 (8.3e-05)	5.4e-06 (2.4e-07)	5.4e-06	0.0033	0.129	0.145
F0.5 90%	Ridge	100,000	0.0184	stratified	8.7e-05 (8.5e-05)	5.6e-06 (2.7e-07)	5.6e-06	0.003	0.129	0.146
F0.5 90%	Ridge	1,000,000	0.0046	naive	-1.6e-06 (1.8e-05)	1.4e-07 (6e-09)	1.4e-07	0.00049	0.0352	0.0377
F0.5 90%	Ridge	1,000,000	0.0046	stratified	-2e-05 (1.8e-05)	1.4e-07 (5.8e-09)	1.4e-07	0.00055	0.0353	0.0375
F0.5 90%	Ridge	1,000,000	0.0092	naive	-4.5e-05 (2.5e-05)	2.1e-07 (9.2e-09)	2e-07	0.00058	0.0704	0.0731
F0.5 90%	Ridge	1,000,000	0.0092	stratified	-5.8e-05 (2.5e-05)	2e-07 (8.8e-09)	2e-07	0.00063	0.0701	0.0731

Table C.64: Bias, MSE, and Variability for $F_{0.5}$ score at the 90th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 95%	Ridge	5,000	0.0046	naive	0.027 (0.00043)	0.00091 (2.8e-05)	0.00019	0.019	0.0409	0.131
F0.5 95%	Ridge	5,000	0.0046	stratified	0.019 (0.00034)	0.00047 (1.5e-05)	0.00011	0.014	0.0421	0.105
F0.5 95%	Ridge	5,000	0.0092	naive	0.013 (0.00057)	0.00048 (2.7e-05)	0.00032	0.025	0.0585	0.224
F0.5 95%	Ridge	5,000	0.0092	stratified	0.0083 (0.00048)	0.0003 (1.4e-05)	0.00023	0.021	0.0538	0.157
F0.5 95%	Ridge	5,000	0.0184	naive	0.0018 (0.0007)	0.00048 (2.2e-05)	0.00048	0.029	0.113	0.269
F0.5 95%	Ridge	5,000	0.0184	stratified	0.0023 (0.00071)	0.00047 (2e-05)	0.00047	0.03	0.115	0.246
F0.5 95%	Ridge	10,000	0.0046	naive	0.0073 (0.00029)	0.00013 (6.5e-06)	8.2e-05	0.012	0.0256	0.0954
F0.5 95%	Ridge	10,000	0.0046	stratified	0.0039 (0.00025)	7.5e-05 (3.4e-06)	6e-05	0.01	0.0331	0.0775
F0.5 95%	Ridge	10,000	0.0092	naive	0.0011 (0.00037)	0.00013 (6e-06)	0.00013	0.016	0.0686	0.152
F0.5 95%	Ridge	10,000	0.0092	stratified	0.00023 (0.00037)	0.00014 (5.9e-06)	0.00014	0.016	0.065	0.134
F0.5 95%	Ridge	10,000	0.0184	naive	-0.0011 (0.00052)	0.00025 (1.1e-05)	0.00025	0.022	0.14	0.237
F0.5 95%	Ridge	10,000	0.0184	stratified	0.00044 (0.00049)	0.00023 (1e-05)	0.00023	0.02	0.144	0.236
F0.5 95%	Ridge	50,000	0.0046	naive	-0.00022 (0.00012)	1.5e-05 (6.7e-07)	1.5e-05	0.0051	0.0383	0.0663
F0.5 95%	Ridge	50,000	0.0046	stratified	-0.0002 (0.00012)	1.3e-05 (6.2e-07)	1.3e-05	0.0052	0.0413	0.0663
F0.5 95%	Ridge	50,000	0.0092	naive	-0.00029 (0.00016)	2.5e-05 (1.1e-06)	2.5e-05	0.0065	0.0871	0.12
F0.5 95%	Ridge	50,000	0.0092	stratified	8.7e-05 (0.00016)	2.5e-05 (1.1e-06)	2.5e-05	0.0067	0.0895	0.121
F0.5 95%	Ridge	50,000	0.0184	naive	-0.00054 (0.00023)	4.6e-05 (2.2e-06)	4.6e-05	0.0085	0.175	0.226
F0.5 95%	Ridge	50,000	0.0184	stratified	-2.4e-05 (0.00022)	4.3e-05 (1.9e-06)	4.3e-05	0.0089	0.18	0.227
F0.5 95%	Ridge	100,000	0.0046	naive	-8.2e-05 (8.9e-05)	7.1e-06 (3.1e-07)	7.1e-06	0.0037	0.0459	0.0637
F0.5 95%	Ridge	100,000	0.0046	stratified	5.4e-05 (9e-05)	7.3e-06 (3.2e-07)	7.3e-06	0.0036	0.0463	0.0631
F0.5 95%	Ridge	100,000	0.0092	naive	-0.00021 (0.00012)	1.3e-05 (5.5e-07)	1.2e-05	0.0048	0.0971	0.119
F0.5 95%	Ridge	100,000	0.0092	stratified	-8.6e-05 (0.00012)	1.2e-05 (5.5e-07)	1.2e-05	0.0045	0.097	0.118
F0.5 95%	Ridge	100,000	0.0184	naive	-0.0002 (0.00016)	2.1e-05 (8.8e-07)	2e-05	0.0061	0.187	0.216
F0.5 95%	Ridge	100,000	0.0184	stratified	1.3e-05 (0.00016)	2.2e-05 (1.1e-06)	2.2e-05	0.0061	0.186	0.225
F0.5 95%	Ridge	1,000,000	0.0046	naive	-5.2e-06 (3.7e-05)	5.7e-07 (2.5e-08)	5.7e-07	0.001	0.055	0.0595
F0.5 95%	Ridge	1,000,000	0.0046	stratified	-3.6e-05 (3.7e-05)	6.1e-07 (2.7e-08)	6.1e-07	0.001	0.0548	0.0596
F0.5 95%	Ridge	1,000,000	0.0092	naive	-7.9e-05 (5e-05)	8.3e-07 (3.6e-08)	8.3e-07	0.0012	0.108	0.113
F0.5 95%	Ridge	1,000,000	0.0092	stratified	-4.4e-05 (4.8e-05)	8.1e-07 (3.7e-08)	8.1e-07	0.0012	0.107	0.113

Table C.65: Bias, MSE, and Variability for $F_{0.5}$ score at the 95th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 99%	Ridge	5,000	0.0046	naive	0.16 (0.0021)	0.029 (0.00082)	0.0047	0.085	0.102	0.568
F0.5 99%	Ridge	5,000	0.0046	stratified	0.15 (0.0021)	0.026 (0.0009)	0.0043	0.073	0.0926	0.833
F0.5 99%	Ridge	5,000	0.0092	naive	0.097 (0.0017)	0.012 (0.00039)	0.003	0.076	0.149	0.517
F0.5 99%	Ridge	5,000	0.0092	stratified	0.095 (0.0017)	0.012 (0.00048)	0.0029	0.071	0.12	0.769
F0.5 99%	Ridge	5,000	0.0184	naive	0.042 (0.0016)	0.0045 (0.00019)	0.0027	0.07	0.205	0.529
F0.5 99%	Ridge	5,000	0.0184	stratified	0.035 (0.0016)	0.004 (0.00018)	0.0027	0.071	0.203	0.538
F0.5 99%	Ridge	10,000	0.0046	naive	0.06 (0.0011)	0.0047 (0.00016)	0.0012	0.044	0.0892	0.322
F0.5 99%	Ridge	10,000	0.0046	stratified	0.054 (0.001)	0.004 (0.00014)	0.0011	0.041	0.0878	0.319
F0.5 99%	Ridge	10,000	0.0092	naive	0.026 (0.0012)	0.0021 (9.4e-05)	0.0014	0.052	0.125	0.344
F0.5 99%	Ridge	10,000	0.0092	stratified	0.023 (0.0012)	0.002 (9.5e-05)	0.0015	0.051	0.119	0.379
F0.5 99%	Ridge	10,000	0.0184	naive	0.0013 (0.0013)	0.0017 (7.8e-05)	0.0017	0.054	0.179	0.459
F0.5 99%	Ridge	10,000	0.0184	stratified	-0.00096 (0.0013)	0.0017 (6.9e-05)	0.0017	0.058	0.206	0.447
F0.5 99%	Ridge	50,000	0.0046	naive	-0.0009 (0.00044)	0.00019 (8.9e-06)	0.00019	0.018	0.0751	0.171
F0.5 99%	Ridge	50,000	0.0046	stratified	-0.0012 (0.00046)	0.00021 (8.6e-06)	0.00021	0.02	0.0776	0.163
F0.5 99%	Ridge	50,000	0.0092	naive	-0.0014 (0.00054)	0.00029 (1.3e-05)	0.00029	0.022	0.156	0.265
F0.5 99%	Ridge	50,000	0.0092	stratified	-0.0003 (0.00057)	0.00032 (1.4e-05)	0.00032	0.024	0.162	0.265
F0.5 99%	Ridge	50,000	0.0184	naive	-0.0021 (0.00057)	0.00033 (1.4e-05)	0.00032	0.025	0.275	0.387
F0.5 99%	Ridge	50,000	0.0184	stratified	-0.0024 (0.00056)	0.00032 (1.5e-05)	0.00031	0.023	0.263	0.387
F0.5 99%	Ridge	100,000	0.0046	naive	-0.00059 (0.00033)	0.0001 (4.3e-06)	1e-04	0.014	0.0945	0.155
F0.5 99%	Ridge	100,000	0.0046	stratified	0.00016 (0.00033)	0.0001 (4.4e-06)	0.0001	0.014	0.0988	0.155
F0.5 99%	Ridge	100,000	0.0092	naive	-0.0015 (0.0004)	0.00015 (6.6e-06)	0.00015	0.017	0.182	0.255
F0.5 99%	Ridge	100,000	0.0092	stratified	-0.00067 (0.0004)	0.00015 (6.6e-06)	0.00015	0.017	0.18	0.258
F0.5 99%	Ridge	100,000	0.0184	naive	-0.0014 (0.00042)	0.00016 (7.3e-06)	0.00016	0.017	0.297	0.378
F0.5 99%	Ridge	100,000	0.0184	stratified	-0.00076 (0.00042)	0.00016 (7.3e-06)	0.00016	0.017	0.288	0.373
F0.5 99%	Ridge	1,000,000	0.0046	naive	-9.1e-05 (0.00015)	9.1e-06 (4.2e-07)	9.1e-06	0.004	0.122	0.141
F0.5 99%	Ridge	1,000,000	0.0046	stratified	-7.4e-05 (0.00014)	9e-06 (3.9e-07)	9e-06	0.0041	0.121	0.141
F0.5 99%	Ridge	1,000,000	0.0092	naive	-0.00029 (0.00016)	9.6e-06 (4.4e-07)	9.5e-06	0.004	0.218	0.238
F0.5 99%	Ridge	1,000,000	0.0092	stratified	-0.00034 (0.00017)	1e-05 (4.5e-07)	9.9e-06	0.0043	0.218	0.237

Table C.66: Bias, MSE, and Variability for $F_{0.5}$ score at the 99th percentile, ridge regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 90%	Logistic	5,000	0.0046	naive	0.0022 (0.00016)	3.6e-05 (2.9e-06)	3.1e-05	0.0054	0.00499	0.0371
F0.5 90%	Logistic	5,000	0.0046	stratified	0.0018 (0.00017)	3.7e-05 (3.1e-06)	3.3e-05	0.0054	0.00487	0.0417
F0.5 90%	Logistic	5,000	0.0092	naive	0.0075 (0.00029)	0.00016 (7.6e-06)	0.0001	0.011	0.0104	0.0931
F0.5 90%	Logistic	5,000	0.0092	stratified	0.0065 (0.00027)	0.00012 (5.2e-06)	8.2e-05	0.01	0.0107	0.0731
F0.5 90%	Logistic	5,000	0.0184	naive	-0.0021 (0.00044)	0.0002 (1.4e-05)	0.00019	0.017	0.0218	0.131
F0.5 90%	Logistic	5,000	0.0184	stratified	-0.0031 (0.00046)	0.00023 (1.6e-05)	0.00022	0.017	0.0219	0.132
F0.5 90%	Logistic	10,000	0.0046	naive	0.0045 (0.00015)	3.9e-05 (1.6e-06)	1.8e-05	0.0057	0.00867	0.0379
F0.5 90%	Logistic	10,000	0.0046	stratified	0.0037 (0.00013)	2.8e-05 (1.2e-06)	1.4e-05	0.005	0.00626	0.0342
F0.5 90%	Logistic	10,000	0.0092	naive	-0.0014 (0.00023)	4.7e-05 (2.1e-06)	4.6e-05	0.009	0.0252	0.0672
F0.5 90%	Logistic	10,000	0.0092	stratified	-0.0018 (0.00022)	4.2e-05 (1.8e-06)	3.9e-05	0.0087	0.0253	0.0657
F0.5 90%	Logistic	10,000	0.0184	naive	-0.0028 (0.00029)	8.3e-05 (3.6e-06)	7.6e-05	0.012	0.0858	0.135
F0.5 90%	Logistic	10,000	0.0184	stratified	-0.003 (0.00029)	8.7e-05 (4.1e-06)	7.8e-05	0.012	0.0756	0.137
F0.5 90%	Logistic	50,000	0.0046	naive	-0.00053 (6.5e-05)	4.1e-06 (1.8e-07)	3.8e-06	0.0025	0.025	0.037
F0.5 90%	Logistic	50,000	0.0046	stratified	-0.00048 (6.7e-05)	4.2e-06 (1.9e-07)	3.9e-06	0.0026	0.0248	0.038
F0.5 90%	Logistic	50,000	0.0092	naive	-0.00048 (9.1e-05)	7.8e-06 (3.5e-07)	7.5e-06	0.0037	0.0588	0.0754
F0.5 90%	Logistic	50,000	0.0092	stratified	-0.00024 (8.6e-05)	6.8e-06 (3e-07)	6.7e-06	0.0036	0.0587	0.0764
F0.5 90%	Logistic	50,000	0.0184	naive	-0.0005 (0.00012)	1.3e-05 (6.1e-07)	1.3e-05	0.005	0.121	0.145
F0.5 90%	Logistic	50,000	0.0184	stratified	-0.00035 (0.00012)	1.3e-05 (5.8e-07)	1.3e-05	0.005	0.123	0.146
F0.5 90%	Logistic	100,000	0.0046	naive	-0.00018 (4.6e-05)	2e-06 (9.6e-08)	2e-06	0.0018	0.0292	0.0391
F0.5 90%	Logistic	100,000	0.0046	stratified	-0.00019 (4.6e-05)	1.9e-06 (8.4e-08)	1.9e-06	0.0018	0.0298	0.0381
F0.5 90%	Logistic	100,000	0.0092	naive	-0.00023 (6.2e-05)	3.3e-06 (1.5e-07)	3.3e-06	0.0024	0.0635	0.0743
F0.5 90%	Logistic	100,000	0.0092	stratified	-0.00018 (6.4e-05)	3.5e-06 (1.6e-07)	3.5e-06	0.0026	0.0615	0.0748
F0.5 90%	Logistic	100,000	0.0184	naive	-0.00016 (8.5e-05)	5.9e-06 (2.5e-07)	5.9e-06	0.0034	0.129	0.143
F0.5 90%	Logistic	100,000	0.0184	stratified	-0.00022 (8.4e-05)	5.6e-06 (2.4e-07)	5.6e-06	0.0034	0.129	0.143
F0.5 90%	Logistic	1,000,000	0.0046	naive	-2.4e-05 (1.8e-05)	1.4e-07 (6.1e-09)	1.4e-07	0.00052	0.0352	0.0376
F0.5 90%	Logistic	1,000,000	0.0046	stratified	-6.9e-06 (1.9e-05)	1.3e-07 (5.6e-09)	1.3e-07	0.00051	0.0354	0.0375
F0.5 90%	Logistic	1,000,000	0.0092	naive	-2.7e-06 (2.6e-05)	2.1e-07 (9.8e-09)	2.1e-07	0.00062	0.0702	0.0732
F0.5 90%	Logistic	1,000,000	0.0092	stratified	3e-06 (2.5e-05)	2.1e-07 (9.4e-09)	2.1e-07	0.00064	0.0699	0.0732

Table C.67: Bias, MSE, and Variability for $F_{0.5}$ score at the 90th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 95%	Logistic	5,000	0.0046	naive	0.0031 (0.0003)	0.00012 (1.1e-05)	0.00011	0.0091	0.00375	0.0714
F0.5 95%	Logistic	5,000	0.0046	stratified	0.0031 (0.00033)	0.00014 (1.2e-05)	0.00013	0.0099	0.00487	0.0748
F0.5 95%	Logistic	5,000	0.0092	naive	0.02 (0.00054)	0.00078 (3e-05)	0.00038	0.022	0.01	0.143
F0.5 95%	Logistic	5,000	0.0092	stratified	0.02 (0.00049)	0.00069 (2.5e-05)	0.00031	0.019	0.0107	0.124
F0.5 95%	Logistic	5,000	0.0184	naive	0.0017 (0.00073)	0.00056 (3.7e-05)	0.00056	0.03	0.0218	0.238
F0.5 95%	Logistic	5,000	0.0184	stratified	-8.9e-05 (0.00073)	0.00059 (3.9e-05)	0.00059	0.028	0.022	0.196
F0.5 95%	Logistic	10,000	0.0046	naive	0.012 (0.00028)	0.00021 (8.3e-06)	6.6e-05	0.011	0.00867	0.0832
F0.5 95%	Logistic	10,000	0.0046	stratified	0.011 (0.00022)	0.00017 (5.7e-06)	4.7e-05	0.0095	0.00647	0.0651
F0.5 95%	Logistic	10,000	0.0092	naive	0.00015 (0.0004)	0.00014 (6.6e-06)	0.00014	0.016	0.0399	0.118
F0.5 95%	Logistic	10,000	0.0092	stratified	-0.0012 (0.00038)	0.00013 (6.1e-06)	0.00013	0.015	0.0321	0.117
F0.5 95%	Logistic	10,000	0.0184	naive	-0.0048 (0.00052)	0.00028 (1.2e-05)	0.00026	0.022	0.111	0.208
F0.5 95%	Logistic	10,000	0.0184	stratified	-0.0046 (0.00052)	0.00028 (1.2e-05)	0.00026	0.022	0.108	0.219
F0.5 95%	Logistic	50,000	0.0046	naive	-0.001 (0.00012)	1.6e-05 (6.9e-07)	1.5e-05	0.0051	0.0352	0.0587
F0.5 95%	Logistic	50,000	0.0046	stratified	-0.00095 (0.00013)	1.6e-05 (7.3e-07)	1.6e-05	0.0052	0.0349	0.0608
F0.5 95%	Logistic	50,000	0.0092	naive	-0.00096 (0.00017)	2.9e-05 (1.4e-06)	2.8e-05	0.0069	0.0836	0.12
F0.5 95%	Logistic	50,000	0.0092	stratified	-0.00051 (0.00017)	2.7e-05 (1.2e-06)	2.6e-05	0.007	0.0852	0.119
F0.5 95%	Logistic	50,000	0.0184	naive	-0.0011 (0.00022)	4.7e-05 (2e-06)	4.6e-05	0.0094	0.177	0.219
F0.5 95%	Logistic	50,000	0.0184	stratified	-0.00054 (0.00023)	4.9e-05 (2.3e-06)	4.9e-05	0.0093	0.175	0.224
F0.5 95%	Logistic	100,000	0.0046	naive	-0.0004 (9.4e-05)	8.1e-06 (3.6e-07)	8e-06	0.0038	0.0448	0.0628
F0.5 95%	Logistic	100,000	0.0046	stratified	-0.00036 (9.2e-05)	7.8e-06 (3.4e-07)	7.6e-06	0.0038	0.0444	0.0632
F0.5 95%	Logistic	100,000	0.0092	naive	-0.00048 (0.00013)	1.4e-05 (6e-07)	1.4e-05	0.0052	0.0946	0.117
F0.5 95%	Logistic	100,000	0.0092	stratified	-0.00039 (0.00013)	1.4e-05 (6.3e-07)	1.4e-05	0.0047	0.0953	0.118
F0.5 95%	Logistic	100,000	0.0184	naive	-0.00024 (0.00016)	2.2e-05 (9.4e-07)	2.2e-05	0.0066	0.189	0.217
F0.5 95%	Logistic	100,000	0.0184	stratified	-0.00033 (0.00016)	2e-05 (8.6e-07)	2e-05	0.0062	0.187	0.213
F0.5 95%	Logistic	1,000,000	0.0046	naive	-2.8e-05 (3.7e-05)	5.8e-07 (2.6e-08)	5.8e-07	0.0011	0.0547	0.06
F0.5 95%	Logistic	1,000,000	0.0046	stratified	-4.2e-06 (3.8e-05)	6e-07 (2.6e-08)	6e-07	0.001	0.055	0.0595
F0.5 95%	Logistic	1,000,000	0.0092	naive	7.1e-06 (5e-05)	8.8e-07 (4.2e-08)	8.9e-07	0.0012	0.108	0.114
F0.5 95%	Logistic	1,000,000	0.0092	stratified	1.3e-05 (5.1e-05)	9e-07 (4e-08)	9e-07	0.0013	0.107	0.114

Table C.68: Bias, MSE, and Variability for $F_{0.5}$ score at the 95th percentile, logistic regression.

							CV value			
Metric	Method	Training set size	Rate	Sampling strategy	Bias (MCE)	MSE (MCE)	Variance	IQR	Min.	Max.
F0.5 99%	Logistic	5,000	0.0046	naive	0.03 (0.0012)	0.0025 (0.00013)	0.0016	0.065	0.0025	0.227
F0.5 99%	Logistic	5,000	0.0046	stratified	0.031 (0.0013)	0.003 (0.0002)	0.0021	0.07	0.00411	0.333
F0.5 99%	Logistic	5,000	0.0092	naive	0.082 (0.0017)	0.01 (0.00044)	0.0037	0.06	0.00749	0.5
F0.5 99%	Logistic	5,000	0.0092	stratified	0.085 (0.0016)	0.01 (0.00034)	0.0032	0.058	0.01	0.4
F0.5 99%	Logistic	5,000	0.0184	naive	0.049 (0.0015)	0.0048 (0.0002)	0.0024	0.06	0.0436	0.396
F0.5 99%	Logistic	5,000	0.0184	stratified	0.046 (0.0015)	0.0046 (0.00019)	0.0025	0.06	0.0223	0.425
F0.5 99%	Logistic	10,000	0.0046	naive	0.07 (0.00095)	0.0058 (0.00016)	0.00091	0.036	0.017	0.25
F0.5 99%	Logistic	10,000	0.0046	stratified	0.071 (0.001)	0.0061 (0.00018)	0.001	0.038	0.00624	0.306
F0.5 99%	Logistic	10,000	0.0092	naive	0.034 (0.001)	0.0022 (9.3e-05)	0.001	0.041	0.0868	0.268
F0.5 99%	Logistic	10,000	0.0092	stratified	0.031 (0.001)	0.002 (8.6e-05)	0.001	0.045	0.0504	0.276
F0.5 99%	Logistic	10,000	0.0184	naive	-0.0023 (0.0012)	0.0014 (6.4e-05)	0.0014	0.051	0.155	0.405
F0.5 99%	Logistic	10,000	0.0184	stratified	-0.0045 (0.0012)	0.0015 (6.5e-05)	0.0015	0.05	0.14	0.373
F0.5 99%	Logistic	50,000	0.0046	naive	-0.0021 (0.00044)	0.0002 (8.2e-06)	0.00019	0.02	0.0602	0.149
F0.5 99%	Logistic	50,000	0.0046	stratified	-0.0022 (0.00043)	0.0002 (8.3e-06)	0.00019	0.02	0.0681	0.158
F0.5 99%	Logistic	50,000	0.0092	naive	-0.0036 (0.00056)	0.00033 (1.4e-05)	0.00032	0.024	0.149	0.257
F0.5 99%	Logistic	50,000	0.0092	stratified	-0.0025 (0.00055)	0.00031 (1.4e-05)	0.00031	0.024	0.147	0.261
F0.5 99%	Logistic	50,000	0.0184	naive	-0.0034 (0.00058)	0.00035 (1.5e-05)	0.00034	0.026	0.255	0.383
F0.5 99%	Logistic	50,000	0.0184	stratified	-0.0026 (0.0006)	0.00036 (1.6e-05)	0.00035	0.026	0.261	0.386
F0.5 99%	Logistic	100,000	0.0046	naive	-0.0015 (0.00033)	0.00011 (4.8e-06)	0.00011	0.014	0.0814	0.154
F0.5 99%	Logistic	100,000	0.0046	stratified	-0.0011 (0.00034)	0.00011 (4.9e-06)	0.00011	0.014	0.0886	0.155
F0.5 99%	Logistic	100,000	0.0092	naive	-0.002 (0.00041)	0.00016 (7.2e-06)	0.00016	0.017	0.177	0.256
F0.5 99%	Logistic	100,000	0.0092	stratified	-0.002 (0.00041)	0.00016 (7.4e-06)	0.00016	0.016	0.17	0.256
F0.5 99%	Logistic	100,000	0.0184	naive	-0.00098 (0.00041)	0.00016 (7.5e-06)	0.00016	0.017	0.283	0.38
F0.5 99%	Logistic	100,000	0.0184	stratified	-0.0015 (0.00041)	0.00015 (7e-06)	0.00015	0.016	0.293	0.37
F0.5 99%	Logistic	1,000,000	0.0046	naive	-0.00013 (0.00014)	9.3e-06 (4.1e-07)	9.3e-06	0.004	0.122	0.141
F0.5 99%	Logistic	1,000,000	0.0046	stratified	-0.00027 (0.00014)	9.1e-06 (4.1e-07)	9e-06	0.004	0.123	0.143
F0.5 99%	Logistic	1,000,000	0.0092	naive	6.8e-06 (0.00017)	1.1e-05 (5.1e-07)	1.1e-05	0.0047	0.218	0.24
F0.5 99%	Logistic	1,000,000	0.0092	stratified	0.00014 (0.00017)	1.1e-05 (4.9e-07)	1.1e-05	0.0044	0.217	0.236

Table C.69: Bias, MSE, and Variability for $F_{0.5}$ score at the 99th percentile, logistic regression.