Technische Universität München
TUM School of Life Sciences

Individual participant data modeling approaches for clinical risk model validation and bias adjustment

Yiyao Chen

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Aurélien Tellier

Prüfende der Dissertation:

1. Prof. Donna P. Ankerst, Ph.D.
2. Prof. Byeongyeob Choi, Ph.D.

Die Dissertation wurde am 24.05.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 30.09.2022 angenommen.

Abstract

After the development of clinical risk prediction models, external validation of the prediction models using data different from the training one is an essential step to assess if the models can perform well with different samples and can be recommended for applications in practice. Usually, the calibration and discrimination of the model are evaluated in the validation. In many clinical trials, the outcome of interest is only known for certain participants who go through the diagnostic verification process but not the rest. Those verified participants are usually not a random sample of the population but rather those who meet the verification criteria. The clinical prediction models are built on the subgroup of verified participants in the training cohort and then, validated with external samples. The validation cohort may differ from the training one in many aspects including the distribution of risk factors, the prevalence of the event of interest, and the verification process. Such heterogeneity between training and validation cohorts could bias the external validation results leading to a fallacious conclusion.

This thesis focuses on accommodating the selection bias coming from different distributions of risk factors and the verification bias coming from different diagnosis test schemes between the training and validation cohorts in the external validation of prediction models, where it requires that the individual participant data from both training and validation cohorts should be available. A novel method is proposed to accommodate the selection and verification biases. The concepts of “reproducibility” and “transportability” is formalized in the discussion of selection bias adjustment.

The proposed framework results in weighted versions of the usual performance metrics with different weights addressing verification bias, selection bias, or the combination of the two. The novel approach is illustrated with a simulation study and a real data example from two large North American prostate cancer screening and prevention trials. The simulation study shows that the weighted metrics could perfectly correct the bias when the unweighted ones are distorted. Researchers are encouraged to share data underlying their published risk prediction models to facilitate external validation. The proposed performance measures are recommended as reference values in external validation of risk prediction models to account for the impact of differences in risk factor distributions or verification processes between training and validation cohorts.

Zusammenfassung

Nach der Entwicklung klinischer Risikovorhersagemodelle ist die externe Validierung der Vorhersagemodelle mit anderen Daten als dem Trainingsmodell ein wesentlicher Schritt, um zu beurteilen, ob die Modelle mit verschiedenen Stichproben gut funktionieren und für Anwendungen in der Praxis empfohlen werden können. Üblicherweise werden bei der Validierung die Kalibrierung und Diskrimination des Modells evaluiert. In vielen klinischen Studien ist das interessierende Ergebnis nur bestimmten Teilnehmern bekannt, die den diagnostischen Verifizierungsprozess durchlaufen, nicht aber den Rest. Diese verifizierten Teilnehmer sind in der Regel keine zufällige Stichprobe der Bevölkerung, sondern diejenigen, die die Verifizierungskriterien erfüllen. Die klinischen Vorhersagemodelle werden auf der Untergruppe der verifizierten Teilnehmer in der Trainingskohorte aufgebaut und dann mit externen Stichproben validiert. Die Validierungskohorte kann sich in vielen Aspekten von der Schulungskohorte unterscheiden, einschließlich der Verteilung von Risikofaktoren, der Prävalenz des interessierenden Ereignisses und des Überprüfungsprozesses. Eine solche Heterogenität zwischen Schulungs- und Validierungskohorten könnte die externen Validierungsergebnisse verfälschen und zu einer falschen Schlussfolgerung führen.

In dieser Arbeit geht es darum, den Selektionsbias, der sich aus der unterschiedlichen Verteilung der Risikofaktoren ergibt, und den Verifikationsbias, der sich aus den unterschiedlichen Diagnosetestschemata zwischen den Trainings- und Validierungskohorten ergibt, bei der externen Validierung von Vorhersagemodellen zu berücksichtigen, was voraussetzt, dass die individuellen Teilnehmerdaten sowohl aus den Trainings- als auch aus den Validierungskohorten verfügbar sind. Es wird eine neuartige Methode vorgeschlagen, um die Auswahl- und Verifizierungsverzerrungen zu berücksichtigen. Die Konzepte der "reproducibility" und "transportability" werden in der Diskussion über die Anpassung der Selektionsverzerrungen formalisiert.

Der vorgeschlagene Rahmen führt zu gewichteten Versionen der üblichen Leistungsmetriken mit unterschiedlichen Gewichtungen, um Verifikationsverzerrungen, Selektionsverzerrungen oder die Kombination der beiden zu berücksichtigen. Der neue Ansatz wird anhand einer Simulationsstudie und eines realen Datenbeispiels aus zwei großen nordamerikanischen Prostatakrebs-Screening- und Präventionsstudien veranschaulicht. Die Simulationsstudie zeigt, dass die gewichteten Metriken die Verzerrung perfekt korrigieren können, wenn die ungewichteten verzerrt sind. Die Forscher werden aufgefordert, die ihren veröffentlichten Risikovorhersagemodellen zugrunde liegenden Daten mitzuteilen, um eine externe Validierung zu erleichtern. Die vorgeschlagenen Leistungsmaße werden als Referenzwerte für

die externe Validierung von Risikovorhersagemodellen empfohlen, um die Auswirkungen von Unterschieden in der Verteilung der Risikofaktoren oder der Verifizierungsprozesse zwischen Trainings- und Validierungskohorten zu berücksichtigen.

Contents

Abstract iii

Zusammenfassung iv

Contents vi

List of Figures viii

List of Tables x

1 Introduction 1

1.1 Risk prediction models 2

1.2 Model validation 5

1.3 Heterogeneity between cohorts 8

1.4 Individual participant data 9

1.5 Outline 12

2 Two large North American prostate cancer screening and prevention trials 13

2.1 Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial . . . 13

2.2 Selenium and Vitamin E Cancer Prevention Trial (SELECT) 14

2.3 Data processing 14

3 Standard validation of risk prediction models 30

3.1 Research context 30

3.2 Notations and metrics 32

3.3 Simple analytic example 41

3.4 Application 44

3.5 Summary 47

4 A novel external validation method for adjusting for training and validation cohort differences 49

4.1 Research context 49

4.2 Method 51

4.3 Simulation study 67

4.4 Prostate cancer example 71

4.5 Discussion 84

5 Summary and outlook 98

Contributing manuscript 104

Acronyms 105

Bibliography 106

Acknowledgements 119

List of Figures

2.1 Univariable odds ratios for being in PLCO versus prevalence of risk factors for the 62874 combined participants of PLCO and SELECT. 19

2.2 Univariable odds ratios for having a biopsy versus the prevalence of dichotomized prostate cancer risk factors in 30245 PLCO and 32629 SELECT participants. 20

2.3 Univariable odds ratios for having prostate cancer versus the prevalence of dichotomized prostate cancer risk factors in 3813 PLCO and 4185 SELECT verified participants. 21

2.4 Density of the time since study registration until prostate-specific antigen (PSA) test, digital rectal exam (DRE) test, or detection of prostate cancer (PCA) in 1833 PLCO and 2028 SELECT participants with prostate cancer. . 22

2.5 Univariable odds ratios for being in PLCO versus prevalence of dichotomized prostate cancer risk factors in 56121 combined participants of PLCO and SELECT. 26

2.6 Univariable odds ratios for having a biopsy within 5 years after baseline versus the prevalence of dichotomized prostate cancer risk factors evaluating the association between the binary risk factor and the outcome of having a biopsy in 29699 PLCO and 26422 SELECT participants. 27

2.7 Univariable odds ratios for having prostate cancer within 5 years after baseline versus the prevalence of dichotomized prostate cancer risk factors evaluating the association between the binary risk factor and the outcome of having cancer in 29699 PLCO and 26422 SELECT participants. 28

2.8 Cumulative incidence curves for time to first biopsy and prostate cancer diagnosis among 29699 PLCO and 26422 SELECT participants. 29

3.1 Data collection processes of a validation and training set. 33

3.2 Estimated prostate cancer risk from the risk model built on 3813 verified PLCO participants with various PSA and DRE values. 46

3.3 True and false positive rates calculated among 4185 verified or 32629 SELECT participants. 47

4.1 Cumulative baseline hazard function from the Cox model for 5-year prostate cancer risk estimated based on 29699 PLCO participants. 75

4.2 Estimated 5-year prostate cancer risk from the risk model built on 29699 PLCO participants for select risk profiles with various ages at baseline, with-/without family history of prostate cancer, and no prior negative biopsy. 76

4.3 Estimated SELECT 5-year verification probability $\hat{P}_0(V = 1|\mathbf{X}^*)$ on the x-axis versus PLCO 5-year verification probability $\hat{P}_1(V = 1|\mathbf{X}^*)$ on the y-axis estimated from logistic models on 26422 SELECT participants. 81

4.4 Histograms of the selection weights w from the logistic model used for calculation of $\widehat{C}_0^W$ for the 26422 SELECT participants according to baseline risk factor categories. 86

4.5 Histograms of the verification weights v from the logistic model used for calculation of $\widehat{C}_0^V$ for the 26422 SELECT participants according to baseline risk factor categories. 87

4.6 Histograms of the combined selection and verification weights used for calculation of $\widehat{C}_0^{WV}$ for the 26422 SELECT participants according to baseline risk factor categories. 88

4.7 Unweighted and weighted receiver operating characteristics curves among 26422 SELECT participants. 89

4.8 Unweighted and weighted true positive rates (TPRs) and false positive rates (FPRs) among 26422 SELECT participants. 89

4.9 Histograms of the selection weights w from the logistic model in Table 4.7 used for calculation of $\widehat{C}_0^W$ using D^C for the 26422 SELECT participants according to baseline risk factor categories. 95

4.10 Histograms of the verification weights v from the logistic model shown in Table 4.8 used for calculation of $\widehat{C}_0^V$ using D^C for the 26422 SELECT participants according to baseline risk factor categories 96

4.11 Histograms of the combined selection and verification weights used for calculation of $\widehat{C}_0^{WV}$ using D^C for the 26422 SELECT participants according to baseline risk factor categories. 97

List of Tables

2.1 Baseline characteristics and verification status of 30245 PLCO and 32629 SELECT participants. 18

2.2 Baseline characteristics verification status within 5 years of 29699 PLCO and 26422 SELECT participants. 25

3.1 Odds ratios for prostate cancer from logistic regression built on 3813 PLCO participants verified by biopsy. 45

3.2 Odds ratios for prostate cancer from logistic regression built on 4185 SELECT participants verified by biopsy. 45

3.3 Odds ratios for prostate cancer from logistic regression including main effect of binary DRE applied to 3813 PLCO participants verified by biopsy. 46

3.4 Estimated CILs and AUCs among 4185 verified or 32629 SELECT participants. 47

4.1 Unweighted (C_0) and weighted (C_0^W) calibration ratios and weighted AUC_0^W for risk model $R(\mathbf{X})$ under different settings. 57

4.2 Distributions of outcome D, model predictor X, and risk factor Z with resulting C_0 and C_0^W. 58

4.3 Simulation results with risk model $R(X)$. 72

4.4 Simulation results with correctly specified risk model, $R(X, Z_1)$. 73

4.5 Hazard ratios and 95% confidence intervals estimated from the Cox model for 5-year prostate cancer risk fit to 29699 PLCO participants using age as the time metric. 75

4.6 Estimated unweighted calibration ratio (C_1) and area under the receiver operating characteristic curve (AUC_1) in 29699 PLCO participants, using D or censoring outcome D^C. 78

4.7 Odds ratios and 95% confidence intervals from the step-wise logistic selection weight model with outcome (1:in PLCO versus 0: in SELECT) applied to the 56121 participants of both studies (29699 from PLCO and 26422 from SELECT). 80

4.8 Odds ratios and 95% confidence intervals (CIs) from logistic models fit to having a biopsy within 5 years fit to 29699 PLCO and 26422 SELECT participants respectively. 81

4.9 Estimated unweighted and weighted calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants. 83

4.10 Odds ratios from the logistic regression model for the probability of being in PLCO without model selection with outcome (1:in PLCO versus 0: in SELECT) applied to the 56121 participants of both studies (29699 from PLCO and 26422 from SELECT). 90

4.11 Characteristics table of the latest prostate-specific antigen (PSA), indicator of the latest PSA greater than 4 ($I(PSA > 4)$), and digital rectal exam (DRE) among 29699 PLCO and 26422 SELECT participants. 90

4.12 Odds ratios estimated from logistic models for 5-year verification risk fit separately to men PLCO and SELECT. 91

4.13 Sensitivity analysis. Estimated calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants using D in the calculation of the measures. 93

4.14 Sensitivity analysis using D^C. Estimated calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants using the censoring weighted outcome, D^C, in the calculation of the measures. 94

1 Introduction

Nowadays many clinical risk prediction models are available on the internet for individuals around the world seeking information concerning screening for disease. These models often provide the probabilities of having diseases, namely the risk of disease, based on the individual characteristics that are related to disease diagnosis. For example, two major online prostate cancer risk calculators, specifically the Prostate Biopsy Collaborative Group (PBCG) Risk Calculator and the European Randomized Study of Screening for Prostate Cancer (ERSPC) Risk Calculator, both provide the risk of having prostate cancer based on age, prostate-specific antigen (PSA), digital rectal exam (DRE), African ancestry, first-degree family history, and prior negative biopsy [SWOP, 2021; Ankerst et al., 2018]. These variables are common risk factors for prostate cancer, such aging, rising in PSA level, having abnormal DRE, African ancestry [National Cancer Institute, 2021c], and a family history of prostate cancer [Kiciński et al., 2011] all known to increase the cancer risk, while having prior negative biopsy can reduce the cancer risk [Thompson et al., 2006; Ankerst et al., 2018; Nordström et al., 2018; Alberts et al., 2019].

Though we observe substantive improvement in diseases diagnosis strategies and therapies, cancer is still the leading cause of death nowadays, in which prostate cancer is the fifth leading cause of cancer death in men in 2020 worldwide [Sung et al., 2021]. Numerous prostate cancer risk prediction tools have been developed in the past decades providing cancer risks based on symptoms of individual, i.e characteristics of the cancer risk factors, such as 127 unique prostate cancer risk prediction models have been detected in the meta-analysis by Louie et al. [2015]. New risk prediction tools are coming out every now and then, incorporating modern cancer diagnosis information from the the improvement of detection technique and advancement in treatment, to give better risk predictions, such as models based on various genomics markers for prostate cancer from different tissues namely urine, prostate biopsy, or radical prostatectomy [Cucchiara et al., 2018]. Today, investigators are yet on the way to upgrade the risk prediction tools and improve the accuracy of cancer risk predictions.

The underlying models of those risk prediction tools are built on data from the past clinical studies and then, validated before publishing online. One can validate the prediction models within the training sets or with external samples, namely the internal or external validation, respectively. Since investigators expect these tools can benefit disease diagnosis for diverse people upon development in the future, the external validation of the model is an essential step during model development, which can ensure the good performance of the model on samples that are not involved in the model training. The external samples are often from different studies at different study sites, involving different ethnicity, and initiating at different time points compared to the model training cohort. Therefore, the characteristics of individual in the external validation cohorts may differ from those in the training cohort, which heterogeneity can cause poor performance of the developed tools upon external samples.

When the individual participant data from the training and validation cohorts are available in the external validation, we can visualize and compare the heterogeneity between cohorts through characteristics tables and figures, such as a figure comparing the odds ratio for high-grade prostate cancer versus the prevalence of risk factors across different cohorts as shown by Ankerst et al. [2018]. We can also use statistical tests in addition to the characteristics tables to evaluate the significance of the difference between cohorts [Ankerst et al., 2018]. On the contrary, if we only have the data from the validation cohort at hand but not from the training in the external validation of public risk prediction tools, direct comparisons of the distributions of characteristics between cohorts are not feasible. Towards this, one can retrain the prediction model on the validation cohort to ensure fair evaluation [Vergouwe et al., 2010]. Due to optimistic bias from using the re-trained model in the validation set, the revised performance measures are proposed as benchmark values to supplement observed metrics.

In the following of this chapter, we first describe the risk modeling methods used in this thesis in Section 1.1, where the mathematical foundations of multivariable logistic regression and Cox regression are recapped. After that, we review the research context about validation of the clinical risk prediction models in Section 1.2 and focus on the evaluation of discrimination and calibration of a model. Then, we discuss the impact of participants heterogeneity between cohorts on external validation and the current approaches to address them regardless having individual participant data or not in Section 1.3 and 1.4, respectively. We close this chapter with an outline of the thesis in Section 1.5.

1.1 Risk prediction models

When estimating the risk of cancer using dichotomous cancer status, i.e. having cancer or not, as the response, we often apply the multivariable logistic regression model with several predictors, which model is a standard method and is frequently used when we have binary response [Riley et al., 2016; Meurer and Tolles, 2017; van Leeuwen et al., 2017; Shipe et al., 2019; Bhat et al., 2019; Steyerberg, 2019]. Other modeling techniques include the penalized logistic models, such as least absolute shrinkage and selection operator (LASSO) regression to exclude trivial predictors [Tibshirani, 1996; Kim et al., 2018; Steyerberg, 2019], and machine learning methods particularly when images are used in the prediction [van der Ploeg et al., 2016; Yala et al., 2019; Mehralivand et al., 2018].

Christodoulou et al. [2019] found that the machine learning methods did not outperform the logistic regressions for clinical prediction modeling in their review of 71 studies and the machine learning algorithm was often criticized to be a ”black-box” lacking necessary clinical transparency [Van Calster et al., 2019]. Since our data examples (See Chapter 2) contains only a few risk factors of prostate cancer nor image, we use the multivariable logistic regression to model the cancer risk for binary cancer status with R software [R Core Team, 2013].

A multivariable logistic regression model assumes that the logit function of the cancer risk, i.e. $ln(p/(1-p))$ with p be the cancer risk, equals to a linear combination of risk factors. Let $\{y_i, x_{i1}, x_{i2}, \dots, x_{ik}\}_{i=1}^n$ be n observations and $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})$. Let y_i equal to 1 for cancer versus 0 otherwise, which is the realization of the binary cancer outcome Y_i following a Bernoulli distribution $Bernoulli(p_i)$. Let p_i be the probability of having cancer given risk factors, i.e. $P(Y_i = 1|\mathbf{x}_i)$. The multivariable logistic regression has the form of

$$logit(p_i) = ln(\frac{p_i}{1-p_i}) = \sum_{j=0}^{k} x_{ij}\beta_j, \quad i = 1, \dots, n. \tag{1.1}$$

Here $ln(\cdot)$ is the natural logarithm and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_j)$ are the intercept and the coefficients for corresponding risk factors. The likelihood function for the above logistic regression is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}.$$

The log likelihood function for the parameters, $l(\beta)$, follows

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \left(\sum_{j=0}^{k} x_{ij}\beta_j \right) - ln\left(1 + e^{\sum_{j=0}^{k} x_{ij}\beta_j}\right).$$

The first derivative of the log likelihood function with respect to $\beta_j, j = 1, \dots, k$ is

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{1}{1 + e^{\sum_{j=0}^{k} x_{ij}\beta_j}} e^{\sum_{j=0}^{k} x_{ij}\beta_j} x_{ij} = \sum_{i=1}^{n} y_i x_{ij} - p_i x_{ij}.$$

The optimal β_j, i.e. its maximum likelihood estimate, is the value that makes the above first derivative be 0.

To investigate the long-term cancer risk, such as the 5-year cancer risk, given the characteristics, we can conduct survival analysis based on time-to-event data, for which the Cox proportional hazard regression is a common method [Riley et al., 2016; Steyerberg, 2019]. In the survival analysis, we compute the cancer risk at certain time point, which is also known as hazard function reflecting the instantaneous probability of having cancer at this time point. Let T denote the survival time, i.e. time to cancer, the hazard function at time t

is

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = -\frac{dS(t)/dt}{S(t)},$$

where $S(t)$ is the survival function at time t equal to $P(T > t)$, i.e. the probability of remaining cancer-free longer than t. The cumulative hazard function at time t is then

$$H(t) = \int_0^t h(u)du = -ln(S(t))$$

with $ln(\cdot)$ be the natural logarithm.

The Cox proportional hazard regression proposed by Cox [1972] examines how the risk factors affect the cancer risk at given time point. Suppose there are n individuals. Let t_i be the realization of the censored survival time T_i for individual i and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ be the realization of the risk factors for this individual. Let δ_i be 1 if individual i has cancer and 0 otherwise. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ be the corresponding coefficients for the p risk factors. The hazard function from Cox proportional hazard regression at time t for individual i is

$$h(t|\mathbf{x}_i) = h_0(t)e^{\mathbf{x}_i'\boldsymbol{\beta}} = h_0(t)e^{\sum_{j=1}^{p} x_{ij}\beta_j}, \ \ i = 1, \ldots, n, \tag{1.2}$$

where $h_0(t)$ is the baseline hazard function at time t.

Suppose there are m observed cancer cases among these n individuals and let $t_{(1)} < t_{(2)} < ... < t_{(m)}$ be the observed distinct time to cancer for these m cases. Let $\mathcal{R}_i$ denote the risk set at time $t_{(i)}$, i.e. $\{j : j = 1, \ldots, n$ and $t_j \geq t_{(i)}\}$. The corresponding likelihood function is defined as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\delta_i e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\sum\limits_{k \in \mathcal{R}_i} e^{\mathbf{x}_k'\boldsymbol{\beta}}}.$$

This likelihood function does not depend on the baseline hazard function, $h_0(t)$, and hence, is known as partial likelihood [Cox, 1975]. The log partial likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[e^{\mathbf{x}_i'\boldsymbol{\beta}} - ln \left(\sum_{k \in \mathcal{R}_i}^{n} e^{\mathbf{x}_k'\boldsymbol{\beta}} \right) \right].$$

Then, the first derivative with respect to β_j is

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i' - \frac{\sum\limits_{k \in \mathcal{R}_i} e^{\mathbf{x}_k'\boldsymbol{\beta}} \mathbf{x}_{kj}}{\sum\limits_{k \in \mathcal{R}_i}^{n} e^{\mathbf{x}_k'\boldsymbol{\beta}}} \right], \ \ j = 1, \ldots, p.$$

Solving $\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j}, \ldots, \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p} \right) = (0, \ldots, 0)$ obtains the maximum partial likelihood estimates of coefficients. The Hessian matrix of $l(\boldsymbol{\beta})$ being negative defined ensures a unique solution for the partial likelihood function [Karim and Islam, 2019].

Once the estimated coefficients of the Cox regression are obtained, we can compute the estimated baseline hazard at distinct cancer time $t_{(i)}$ as

$$\widehat{h_0}(t_{(i)}) = \frac{d_i}{\sum\limits_{k \in \mathcal{R}_i} e^{\mathbf{x}'_k \widehat{\boldsymbol{\beta}}}}, \quad i = 1, \ldots, m,$$

where d_i is the number of cancer cases at the distinct cancer time $t_{(i)}$ [Moore, 2016; Breslow, 1972, 1975]. In this thesis, we implement the Cox proportional hazard regression model by the `survival` package in R software [Therneau, 2021; R Core Team, 2013].

1.2 Model validation

Once a risk prediction model has been developed, we can evaluate it internally with its deriving data as well as externally with other new samples to assess the accuracy of its predictions. Internal validation refers to validating the model with the training data used to develop the model. Except for simply validating with its original training data set, other common approaches used for internal validation include cross-validation and bootstrapping [Ramspek et al., 2021; Cowley et al., 2019; Steyerberg, 2019]. The former is to split the model development data into training subset for building the model and the remaining for test, while the latter is to generate new data sets out of the model development data via sampling with replacement and evaluate the model upon these newly-generated data sets [Ramspek et al., 2021; Steyerberg and Harrell Jr, 2016]. The internal validation shows the optimistic performance that the prediction model can have. It evaluates rather the sampling variability than the population differences, which the latter can be addressed in the external validation [Cowley et al., 2019]. Hence, one should always perform external validation when building new risk prediction models because the internal validation can never substitute the external one [Moons et al., 2012; Steyerberg et al., 2019].

External validation of a clinical risk prediction model relies on external data samples that are not used in the model training. In the external validation, we can assess the quality of the model with samples similar or different to the training population, namely reproducibility or transportability. The reproducibility reflects the internal validity of the established risk model, while the transportability, also known as generalizability, shows whether the model can perform well upon different but compatible external samples [Steyerberg, 2019; Debray et al., 2015]. Here, the compatible external samples refer to those who are plausible related to the training population, i.e. external samples can be viewed as coming from the same super-population as the training set and the model is reasonable to be applied to these samples [Steyerberg et al., 2001].

Usually, two aspects of model performance are evaluated: the ability to correctly distinguish cases versus non-cases, i.e. the discriminability, and the degree of agreement between observed outcomes and predictions from the model, i.e. calibration [Steyerberg, 2019; Ramspek et al., 2021]. There are other validation assessment aspects including the overall fitness that quantifies the distance between the observations and predictions using the metrics like the explained variation, and the clinical usefulness that evaluates whether the model brings benefit for clinical decision making [Steyerberg et al., 2010; Steyerberg and Vergouwe, 2014]. In this thesis, we focus on assessing the discrimination and calibration of clinical risk prediction models.

Discrimination

The discrimination ability of a clinical risk prediction model refers to whether it can correctly assign the observed cases with higher diseases risk and lower risk to the non-cases. The area-under-the-receiver-operating-characteristic curve (AUC) is often used for evaluating the discrimination, which is calculated as the area under the receiver operating characteristic (ROC) curve that has the true positive rate on the y-axis and false positive rate on the x-axis for binary outcomes [Riley et al., 2016; Ramspek et al., 2021]. The closer the AUC value to one, the better the ability of discrimination of the model. The true positive rate and true negative rate, i.e. 1-false positive rate, are sometimes reported together with clinical models [Simon et al., 2018]. Though these two rates are crucial components in the calculation of AUC, they actually measure the proportions of being correctly classified, whereas AUC focuses on the ability of correctly separating [Steyerberg and Vergouwe, 2014]. They assess different aspects of a prediction model and should not be used interchangeably.

When it comes to survival analysis, Harrell et al. [1982] proposed a concordance index (c-index) to evaluate the discrimination of a model based on time-to-event data, which c-index is the same as the AUC if the outcomes are dichotomous. Because such c-index is based on the order of survival times and predicted risks such that the longer survival time should correspond to lower disease risk, it is affected by the censoring mechanism [Gönen and Heller, 2005; Uno et al., 2011; Steyerberg, 2019]. Several extensions of the Harrell's c-index have been proposed for handling the censoring including the inverse probability of censoring weighting c-index [Uno et al., 2011], involving pre-specified censoring time point in the concordance comparison [Heagerty and Zheng, 2005], and c-index based on the model or linear predictor, i.e. the sum of the risk factors of individual times corresponding coefficients [van Klaveren et al., 2016; Gönen and Heller, 2005]. Royston and Sauerbrei's D statistic for assessing the discrimination of model with survival data is the coefficient from a Cox model regressing the survival outcomes in the validation sample on the scaled rankit of the linear predictors given by the model to be validated as the only predictor [Royston and Sauerbrei, 2004]. This D statistic ranges from 0 to infinity and is independent to censoring given the model is correctly specified [Royston and Sauerbrei, 2004; Rahman et al., 2017]. Logarithm of the D statistics represents the hazard ratio between two equal-size groups with low versus high predicted risk and hence, measures the separation of the survival curves of the two groups. Therefore, a larger value of D statistic stands for better separation and better discrimination of the model [Royston and Sauerbrei, 2004; Royston and Altman, 2013; Austin et al., 2017].

Though widely used, the AUC has some limitations. For example, it ignores the calibration of the model, such that a model lack of fitness could have a good discrimination performance [Hosmer Jr et al., 2013; Lobo et al., 2008; Pencina and D'Agostino, 2015]. The AUC is not sensitive to changes in the predicted risk values, such as adding new predictors into the model or the structure error appears in measurement that the values for all participants vary in the same magnitude, as long as the rank of the risks is preserved [Lobo et al., 2008; Ferri et al., 2005; Pencina et al., 2008; Pajouheshnia et al., 2019]. Researchers also criticize that the AUC treats the falsely predicted positive and negative equally, which the latter is more harmful because it shows a case to be non-case and could hinder the timing of treatment [Lobo et al., 2008].

Other than AUC or c-index, we can also use discrimination slope calculated as the difference in the average predicted risk between the cases and non-cases to evaluate the discrimination with a higher value indicating a better separation [Steyerberg et al., 2010; Pencina and D'Agostino, 2015]. Instead of plotting the ROC curve, Pepe and Janes [2013] examined the discrimination by plotting both true positive and false positive rates, on the y-axis, versus the risk threshold on the x-axis in one figure. Such plot can clearly show the proportions of high-risk participants who are of the interest of researchers and often recommended to be treated, among cases and non-cases. Two curves dispersing from each other a lot represents good discrimination of the risk prediction model.

Calibration

The evaluation of the calibration of a model is to check if the average predicted risk and the case prevalence are identical, which can be measured overall or in subgroups [Stevens and Poppe, 2020; Miller et al., 1991]. Intuitively, we can visualize the calibration of a model by plotting the observed outcomes versus the predicted risks to check if this curve lies on the diagonal. Such plot is known as the calibration plot and is widely used in clinical risk prediction model assessment [Riley et al., 2016; Ankerst et al., 2018; Van Calster et al., 2019].

Quantitatively, in a regression using the observed outcomes as the response and the predicted risks as the only predictor, the coefficient of the predictor when setting the intercept term to be 0 is the calibration slope, whose value close to 1 shows better calibration [Stevens and Poppe, 2020; Steyerberg, 2019]. The calibration-in-the-large (CIL) is the average difference between the predicted risk and observed outcomes measuring the calibration of the model over the whole validation sample [Steyerberg, 2019]. A CIL close to its ideal value 0 indicates less difference between predictions and observations and hence, the model is well-calibrated. Considering the regression used for calculating the calibration slope, the CIL is just the intercept of the regression when the coefficient of the predictor is 1. The Brier score in the form of the average squared difference between the observations and predictions is similarly defined as the CIL. It is also known as the mean squared prediction error measuring the variance between observations and predictions and often used to assess the overall model performance [Assel et al., 2017; Steyerberg et al., 2010]. Other CIL-like metrics for calibration assessment include the absolute error averaging the absolute difference between observations and predictions.

We can also use statistical tests to evaluate the calibration, such as the Hosmer-Lemeshow (HL) test for binary outcomes. It is originally designed to evaluate the goodness-of-fit of the logistic model but is often used for calibration as well [Hosmer and Lemesbow, 1980; Steyerberg, 2019]. The statistic used in the HL test is based on the squared differences between observations and predictions in subgroups of the validation sample and hence, can measure the degree of agreement between them. Since it is based on grouped individuals, HL test is affected by the choice of group partition scheme [Hosmer et al., 1997]. One can also apply statistical tests for the calibration slope and CIL in the context of regression to assess the calibration via the usual tests for the significance of regression coefficients, which the null hypotheses are calibration slope be 1 and CIL be 0, respectively [Steyerberg, 2019]. Both HL and regression-coefficient-based tests are suffered from poor performance when sample sizes are small [Steyerberg, 2019; Ramspek et al., 2021].

In addition to measuring the average gap between observations and predictions, we can also quantify the model validation via the ratio of observations over predictions, i.e. the observed to expected ratio (O/E) [Riley et al., 2016; Debray et al., 2017; Haga et al., 2018; Snell et al., 2021]. Since such a ratio is often used in the disease research to compare the observed disease incidence number versus the expected amount from epidemiology data, it is also known as standardized incidence ratio with an ideal value 1 [Crowson et al., 2016].

1.3 Heterogeneity between cohorts

In a clinical study, the disease status of participants is known only after verifying through gold standard approaches, like prostate cancer confirmation via prostate biopsy. The clinical risk prediction model is then developed on these verified individuals with known disease outcomes. However, the verified individuals within the cohort used to develop the risk prediction tool often differ from the unverified participants from the same cohort in substantive ways. As an example, the online PBCG Risk Calculator (PBCG-RC) for biopsy-detectable prostate cancer has been constructed using individual patient data collected between 2006 to 2017 from ten heterogeneous urological centers in North America and Europe [Ankerst et al., 2018; Tolksdorf et al., 2019]. Participants used for constructing the calculator are not a random sample of men visiting the clinics, but are more commonly presenting with risk factors for prostate cancer, like elevated PSA or abnormal DRE, let alone the referral criteria for prostate biopsy differ across the centers.

Though only based on a subgroup of individuals who have positive disease indication, the developed risk prediction models are usually applied to external samples with temporal or geographical differences compared to the training sample. The external samples and the model training cohort can have large differences in the distributions of individuals characteristics, which hinder the model from making reliable risk predictions. For instance, in a comparison experiment, Carbunaru et al. [2019] applied the PBCG-RC to a sample collected between 2009 to 2014 from five hospitals in Chicago that had a larger amount of Blacks compared to PBCG-RC training cohort (48.5% versus 13%). They agreed with the finding in Ankerst et al. [2018] that the PBCG-RC works well in the Whites, but over-predicts the prostate cancer risks for the Blacks.

In the external validation, researchers have long been aware of the impact of the heterogeneity in participants characteristics, also known as the case-mix difference [Steyerberg, 2019], upon the validation results. Here, the case-mix difference refers to not only the difference in the distributions of risk factors but also in the outcome distribution [Steyerberg, 2019]. Verbeek et al. [2019], Drost et al. [2019], Chen et al. [2021], and Stojadinovic et al. [2020] all claimed the heterogeneity between validation and training cohorts impairs the accuracy of the predicted risks to certain degree in their external validations of prostate cancer risk calculators. The prediction tools work poorly with samples that are drastically different from the training cohorts. The variation in the verification mechanism is another source of cohort heterogeneity that may impact the performance of external validation because the observed disease outcome depends on the verification decision, such as shown by Drost et al. [2019] the difference in the frequency of biopsies between cohorts would bias the predictions given by a prostate cancer active surveillance risk calculator to the external cohort. Overall, the variation in the model performance may attribute to the differences in the distributions of risk factors, verification mechanism, risk factors measurement strategies, and the design of studies [Chen et al., 2021; Drost et al., 2019; Luijken et al., 2019; Ban et al., 2016].

1.4 Individual participant data

In the attempt to address the differences between training and validation cohorts in the distributions of participant characteristics, the availability of individual patient-level data from both the training and validation sets allows the comparisons in terms of risk factor distributions and outcome prevalence between training and external validation cohort. Characteristics tables together with appropriate statistical tests for distributions difference are common practice [Ankerst et al., 2018; Drost et al., 2019; Chen et al., 2021]. Distribution plot of risk factor is an intuitive way for cohorts comparison, such as the graphical displays of cohort-specific risk factor distributions and univariate associations, eliciting transparency in multi-cohort modeling and validation [Ankerst et al., 2018; Tolksdorf et al., 2019].

Utilizing data from both training and validation cohorts, Debray et al. [2015] applied a logistic model with the cohort indicator as the response to estimate the likelihood of being in the training versus validation cohorts, i.e. the membership model. The two cohorts are similar in the distributions of predictors and outcome events if the logistic model distinguishes poorly, such as having an AUC closed to 0.5. Austin et al. [2016] later applied such approach to assess the temporal case-mix difference with data from two different time periods and Steyerberg et al. [2019] used it to check the heterogeneity across cohorts in a meta-analysis. Wang and Lee [2015] permuted individuals from both training and validation cohorts to create a new training and validation pair. They then retrained the model with the permuted training sample and validated the retrained model with the permuted validation sample. They repeated this process multiple times to obtain a permutation $p-$value under the null hypothesis of case-mix similarity with $p < 0.05$ reflecting the existence of the significant difference between cohorts. However, Nieboer et al. [2016] later showed that such permutation method can give misleading conclusions, such as showing homogeneity between cohorts when risk factors distributions are similar but the true predictor effects for risk factors are actually different. Debray et al. [2015] detected the severity of difference in the distributions of risk factors between validation and training cohorts using the individual linear predictor (LP), i.e. the sum of risk factors times their corresponding coefficients, where a larger difference in the mean LP between cohorts reveals greater between-cohort heterogeneity in the distributions of risk factors occurs. Song et al. [2020] applied an adjusted maximum mean discrepancy metric to explain the variation in the performance of the trained model versus the retrained one fit to the validation data in the belief that the variation came from heterogeneity in predictors. Their metric is calculated with the predictors data from both training and validation cohorts. As indicated by its name, the higher the value of the adjusted maximum mean discrepancy, the larger the discrepancy in predictor distributions between training and validation sets.

All the aforementioned discussions are about revealing or evaluating the extent of heterogeneity between cohorts. To eliminating the heterogeneity in risk factor distributions, Powers et al. [2019] matched the risk factor distributions in the validation to the target population by applying a weight to the validation cohort computed as the ratio of the prevalence of risk factor in the target population divided by that in the validation cohort. We can also use propensity score to harmonize the variation in the distribution of participant risk factors between cohorts, which score refers to the conditional probability of being in a certain cohort versus the other given the risk factors and is often estimated with logistic regression built on the combined data with the binary cohort indicator as the response [Rosenbaum and Rubin, 1983]. Weighting a sample by functions of the propensity score to approximate the target population is a conventional approach used in the transportation of causal inference results from one clinical trial to the target cohort [Dahabreh et al., 2019; Westreich et al., 2017; Kern et al., 2016], or resemble a non-random sample towards the general population in survey research [Elliot, 2013; Schonlau et al., 2017; Elliott et al., 2017]. Depending on the target population and the goal of analysis, one can use different forms of the weighting function to harmonize the risk factor distributions between sample and target. Ackerman et al. [2019] weighted the validation sample with a function of propensity score e of being in the training set to resemble the training population, where the weight is calculated as $e/(1-e)$ so that a participant who represents the training cohort better will be up-weighted. Other than weighting individuals from a sample to resemble the target, matching individuals from sample to target based on their propensity scores can also remedy the difference in the distributions of risk factors between cohorts. For instance, we can one-to-one match the individuals who have similar propensity scores between the sample and the target to obtain a new matched sample that has relative identical characteristics distribution compared to the target, where the nearest neighbor can be used to reflect the degree of similarity between individuals [Austin, 2011a,b; Austin and Stuart, 2017].

When there is no access to the individual participant data from the training cohort, the direct comparison of individual characteristics between cohorts is not feasible. To separate the impact of both heterogeneity in characteristics distributions and difference in predictor effect on the external validation results, Vergouwe et al. [2010] introduced two discrimination benchmark values using only the validation cohort. One is a case-mix-corrected $c-$index based on newly-generated disease outcome simulated from the predicted risks. The amount of decreasing in such case-mix-corrected $c-$index compared to the observed $c-$index, i.e. based on the observed outcome, reflects the upper limit of degree of deterioration in model fitness in the validation cohort when the case-mix differences between cohorts are ignored [Nieboer et al., 2016; Austin et al., 2016]. Another benchmark value is a $c-$index by refitting the same model on the validation data and computing the $c-$index based on the refitted model in the validation cohort afterward. The $c-$index from the refitted model can show the best performance this model could achieve with this validation cohort [Vergouwe et al., 2010]. van Klaveren et al. [2016] replaced the observed outcomes of the external validation cohort in the calculation of AUC with the estimated linear predictors from the model built on training cohort to address the impact of case-mix difference on the observed AUC, assuming that the model is perfect for the validation cohort. The difference in the estimated values of their model-based AUC on the training versus validation cohorts quantifies the change of the discrimination ability of the model attributing to the heterogeneity between the two cohorts. Because the observed outcomes from the validation cohorts are not involved in the calculation, the proposed model-based AUC gives the expected discrimination the model could have on the validation and is similar to the case-mix-corrected c-index proposed by Vergouwe et al. [2010] serving as a benchmark value in addition to the observed metric. Royston and Altman [2013] presented the external validation of prognostic model under different disclosure levels of training cohort information, where the availability of individual participant data of training cohort is not required. To carry out a valid evaluation of a model, investigators should at least know its coefficients. Knowing the Kaplan-Meier curves for risk groups and baseline survival function of the training cohort additionally can enable one to assess the calibration of the model, where if the baseline survival functions from the training set are usually unavailable or only available for specific time point, their approximations can be used [Crowson et al., 2016].

1.5 Outline

Motivated by the need to address the impact of heterogeneity in risk factors distributions on the external validation of risk prediction model, we concentrate on accommodating the heterogeneity in risk factors distributions directly in the calculation of performance measures in external validation of risk prediction model when the individual participant data from both training and validation cohorts are accessible in this thesis. We explore the methods to remedy the impact of heterogeneity utilizing the data from two North American prostate cancer screening and prevention trials.

In the following, we first introduce the two prostate cancer trials used throughout this thesis in Chapter 2. After that, we apply the standard external validation method in Chapter 3, where we develop and validate a multivariable logistic regression for prostate cancer risk. In Chapter 4, we accommodate the heterogeneity in the characteristics distributions and verification mechanism between the training and validation cohorts from a weighting point of view and illustrate the proposed novel method using survival data for prostate cancer, where we build a Cox model to estimate the long-term cancer risk. Finally, we summarise our approach and outlook the future works in Chapter 5.

2 Two large North American prostate cancer screening and prevention trials

Throughout this thesis, data from two large North American prostate cancer screening and prevention trials are used, which are the Selenium and Vitamin E Cancer Prevention Trial (SELECT; NCT00006392) and the Prostate, Lung, Colorectal, and Ovarian (PLCO; NCT00002540) Cancer Screening Trial. Several prostate cancer risk models are built and validated in the thesis, for which PLCO is used as the training set, while SELECT is the validation set in all explorations.

2.1 Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial

PLCO was a randomized, controlled trial of screening tests for prostate(P), lung(L), colorectal (C), and ovarian(O) cancers. The trial was designed and sponsored by the National Cancer Institute (NCI) to determine whether screening reduced the mortality of the four kinds of cancer. 76685 male participants were enrolled between November 1993 and July 2001 in 10 centers across the United States. Eligible participants had to be at least 55 years old and up to 74 years old at enrollment and had no prostate cancer before study entry. Participants were assigned to either control or screening arms randomly, resulting in 38345 men in the control arm versus 38340 men in the screening arm [Andriole et al., 2012].

Enrolled participants in the control arm were followed for 13 years and received the standard medical care including occasionally screening, while participants in the intervention arm received annual screening on the prostate-specific antigen (PSA) level in the first 6 years and digital rectal exam (DRE) in the first 4 years after enrollment and were followed by at least 7 additional years [National Cancer Institute, 2021a]. The primary endpoint of the trial was cause-specific mortality of the PLCO cancers [Prorok et al., 2000]. During the trial, participants with PSA greater than 4 nanograms (ng)/milliliter (mL) and/or suspicious DRE results were encouraged to have a diagnostic verification by their physicians [Andriole et al., 2012]. There was no centralized pathological confirmation for cancer, but the pathology labs near the screening center assessed the medical records of men with confirmed prostate cancer and assigned their respective Gleason scores [National Cancer Institute, 2021b; Pinsky et al., 2007].

2.2 Selenium and Vitamin E Cancer Prevention Trial (SELECT)

SELECT was a phase III, randomized, placebo-controlled, four-arm trial to investigate whether selenium (200 micrograms (μg)/day from L-selenomethionine) or vitamin E (400 international unit (IU)/day of all rac-α-tocopheryl acetate) or the combination of both would bring benefit to prostate cancer prevention during a study period of minimum 7 and maximum 12 years. A total of 35533 men were enrolled and randomized between August 22, 2001 and June 24, 2004 from 427 study sites in the United States, Canada and Puerto Rico [Lippman et al., 2005; Klein et al., 2011].

Eligibility criteria were 55 years or older for non-African American men but 50 years or older for African American men due to their increased prostate cancer risk. Further enrollment criteria included a PSA level $\leq$ 4 ng/mL, a normal DRE, and no prior prostate cancer diagnosis before study recruitment [Lippman et al., 2009].

Participants were monitored every 6 months for the development of prostate cancer during the study at their study sites, during which participants were asked to report any new events, including cardiac incidents, diabetes, and any other severe issues not attributable to the study supplements. SELECT recommended to biopsy participants with PSA greater than 4 ng/mL and/or abnormal DRE [Southwest Oncology Group, 2021]. Those who were diagnosed with prostate cancer were followed annually thereafter for their cancer treatments and to update the study endpoints. The primary endpoint of the study was prostate cancer diagnosis determined by routine clinical management and reported to the study sites by participants themselves. The prostate tissue samples and the pathology reports from the participants with suspicious prostate cancer were sent to the central pathology laboratory for confirmation with their corresponding Gleason scores being assigned by the laboratory [Klein et al., 2000, 2011; Lippman et al., 2009]. The study closed early in 2008 after a median follow-up of 5.5 years due to lack of prevention effect [Lippman et al., 2009].

2.3 Data processing

Scenarios with either binary endpoint, i.e. with or without prostate cancer, or survival endpoint, i.e. time to prostate cancer, are explored in this thesis, for which two different pairs of PLCO-SELECT cohorts are created for the analyses with binary or survival endpoints, respectively. For either pair, we use only those participants enrolled in the screening arm of PLCO since several PSA and DRE values from annual screening ensure the availability of enough data. The detailed cohorts construction criteria for each of the two pairs are described in this section.

In PLCO and SELECT, the collected risk factors are not all the same due to different study perspectives. Among these risk factors, we select six of them: age, PSA, DRE, first-degree family history of prostate cancer, prior negative biopsy, and African ancestry as candidate risk factors in the modeling of the prostate cancer risk, which are the common risk factors collected by both cohorts and are largely used in several major online prostate cancer risk calculators [Ankerst et al., 2018; Zaytoun et al., 2011; SWOP, 2021].

Scenario 1

In the analysis for prostate cancer prevalence with binary endpoints, we exclude men who developed any cancer before study entry and did not have any PSA records during the study from both PLCO and SELECT. Among the rest, we set the baseline for men without a biopsy at the age when their latest PSA values were recorded during the study. Their most recent PSA and DRE values recorded during the study observation period are used as the corresponding baseline values. For men with at least one biopsy, we set the baseline at the age of the latest biopsy date for men without cancer or at the age of cancer diagnosis for men with cancer. Their most recent PSA and DRE values recorded within two years before the latest biopsy date or before the cancer diagnosis date are used as the baseline values. We exclude men without eligible PSA and DRE values and with any missing values in first-degree family history, prior negative biopsy, and African ancestry from both cohorts. At last, men at the age $<$ 55 years at baseline are excluded as well, resulting in 30245 PLCO and 32629 SELECT participants for further analysis.

According to the protocols, eligible men in PLCO had to be between 55 to 74 years old but no upper age limit for men in SELECT at study entry. Though we use the age at the most recent PSA recorded date or at the latest biopsy date as the baseline, the inclusion criteria still affect the new baseline age of the selected participants, which the SELECT has a higher percentage of men older than 75 years of age compared to PLCO as shown in Table 2.1. PLCO has higher proportions of participants with PSA greater than 10 ng/mL and abnormal DRE at baseline compared to SELECT (1.6% versus 0.4 % for PSA $>$ 10 ng/mL; 10.1% versus 4.0% for abnormal DRE), which the inclusion criteria of the two studies could be a reason since PLCO does not require the PSA at study entry to be $\leq$ 4 ng/mL. On the other hand, from Table 2.1, SELECT has roughly twice the percentages of participants who have African ancestry, first-degree family history, or prior negative biopsy versus PLCO, while the proportions of being verified are similar between the two cohorts (12.6% versus 12.8%).

In Figure 2.1, we combine the participants from both studies and calculate the odds ratio for being in PLCO in terms of six dichotomized candidate risk factors of prostate cancer.

Men with abnormal DRE have the odds of being in PLCO nearly three times as large as those with normal DRE, while men with PSA $>$ 4 ng/mL are also more likely to be in PLCO with the odds of being in PLCO around 50% higher than those with lower PSA values. Here the odds of being in PLCO refers to the probability of being in PLCO divided by the probability of being in SELECT. The higher the odds of certain event, the higher the probability of the occurrence of such event. Men with abnormal DRE having a higher chance to be in PLCO may be due to that man with PSA $>$ 4 ng/mL had been enrolled in PLCO, but not in SELECT. On the other hand, African Americans have the odds of being in SELECT nearly four times higher compared to non-African Americans. While both trials were conducted in sites throughout North America, the higher odds of being in SELECT for African Americans could be attributed to a special minority recruitment incentive [Cook et al., 2005, 2010]. Men with prior negative biopsy, at the age $>$ 75 years of age, and with a family history of prostate cancer are more likely of being in SELECT than in PLCO with the odds ratios around 0.49, 0.45, and 0.40, respectively.

Since both PLCO and SELECT are recommended to biopsy men with PSA $>$ 4 ng/mL and/or abnormal DRE, Figure 2.2 shows high odds ratios for being biopsy, i.e. being verified, for both situations. However, a PLCO participant with PSA $>$ 4 ng/mL has around twice the odds of receiving a biopsy compared to a SELECT participant (odds ratios: 49.2 versus 24.4), whereas the reverse is true for a participant with abnormal DRE (odds ratios: 12.9 versus 30.8). Odds ratios for the other risk factors do not differ much between the two cohorts. When participants are at an age greater than 75 years, the odds ratios are smaller than one (odds ratios: 0.46 and 0.59) indicating less likely to be verified in both cohorts. This may due to that the elderly only consisted of a small proportion of participants in both cohorts, i.e. 8.3% in PLCO and 16.8% in SELECT as shown in Table 2.1, so that we lack well representative samples of them. Physicians may have the considerations for balancing the harms and benefits of verifying the elderly, which could also confound the outcome [Kotwal and Schonberg, 2017].

PLCO men with African ancestry have higher odds of being verified compared to those without (odds ratio: 1.3), whereas having African ancestry does not affect the odds of biopsy much in SELECT with the odds ratio being around one (odds ratio: 0.9). Having a prior negative biopsy and a first-degree family history of prostate cancer are more likely to be verified in both cohorts. However, men with a family history have a slightly higher odds ratio of being biopsied in SELECT compared to that in PLCO (odds ratios: 1.7 versus 1.5), while the odds of biopsy for men with prior negative biopsy in PLCO is around twice higher than in SELECT (odds ratios: 3.1 versus 1.7). Among the six risk factors, only the age $>$ 75 years decreases the probability of being biopsy in both cohorts, while all the rest increase the likelihood of being verified. For each risk factor, the odds ratios from both cohorts are either both greater than one or smaller except for the African Americans that SELECT has an odds ratio slightly lower than one.

Figure 2.3 depicts the odds ratio of having prostate cancer per binary risk factor versus the prevalence of the respective risk factor among verified participants in PLCO and SELECT, respectively. Having PSA $>$ 4 ng/mL, with a family history of cancer, and having African ancestry all increase the risk of having prostate cancer, i.e. with the odds ratio greater than one, while having a prior negative biopsy decrease the prostate cancer risk in both trials. Verified men in SELECT with abnormal DRE have a higher risk of having prostate cancer compared to those with normal DRE results, while the reverse is true in PLCO that verified men with normal DRE have a higher risk of having cancer (1.4 in SELECT versus 0.5 in PLCO). SELECT excluded men with abnormal DRE from participating in the study but PLCO did not, which explains the higher proportion of having abnormal DRE in PLCO in contrast to that in SELECT. To recall, we use the latest DRE and PSA values taken within two years before the latest biopsy in this scenario. The abnormal DRE related to decreasing prostate cancer risk in PLCO may be due to that the DRE test results were taken much earlier than the detection of prostate cancer in PLCO.

From the left panel of Figure 2.4, we can see that the density curve for time to PSA and the one for time to prostate cancer extend to year 6 after study enrollment in PLCO, while the one for time to DRE stops 4 years after study enrollment. In SELECT, the density curves for time to PSA, DRE, or prostate cancer are roughly overlapped as shown in the right panel of Figure 2.4, which indicates that the timing of PSA test, DRE test, and the detection of prostate cancer are close to each other. Hence, in PLCO, there is a longer time gap between the time of DRE test and the time of prostate cancer compared to that in SELECT, which could detriment the predictive ability of the DRE test results for prostate cancer since the DRE test has been taken too early prior than cancer diagnosed. Other researchers also found that in PLCO, the DRE results did not benefit much for the prostate cancer detection when the PSA level was $\leq$ 4 ng/mL, where the abnormal DRE results only captured 2% of prostate cancer in such situation [Cui et al., 2016]. Moreover, the DRE test results depend on the experience of the physicians and the criteria of “abnormal” resulting in substantially heterogeneous test results across study sites, which variation reduces the accuracy of predicting prostate cancer with solely DRE results [Naji et al., 2018].

Having an age greater than 75 years does not affect the prostate cancer risk much in PLCO, while only slightly increases the cancer risk in SELECT (odds ratios: 0.97 in PLCO versus 1.26 in SELECT in Figure 2.3), which may attribute to that the elderly are rarely joined in clinical trials in general and PLCO had excluded men older than 74 years of age at enrollment. Similarly as in Figure 2.2, the odds ratio of having prostate cancer for men with PSA $>$ 4 ng/mL in PLCO is higher than that in SELECT (odds ratios: 2.3 versus 3.0), while the one for men with abnormal DRE is higher in SELECT (odds ratios: 1.4 versus 0.5). For men older than 75 years of age or having a family history, the corresponding odds ratios of having cancer in SELECT are higher than those in PLCO, while the odds ratios of having cancer for African Americans and for men with prior negative biopsy are higher in

PLCO. Except for abnormal DRE and age $>$ 75 years, the odds ratios from both cohorts for each risk factor are either above one for both or below.

Table 2.1: Baseline characteristics and verification status of 30245 PLCO and 32629 SELECT participants. All p-values from Wilcoxon (Chi-square) tests for numerical (categorical) variables are significant and less than 0.001 except for Prostate-specific antigen (p-value = 0.88) and Verified (p-value = 0.42). Q_1 = the first quartile, Q_3 = the third quartile.

	PLCO (n = 30245)	SELECT (n = 32629)
Age (year)		
(min., Q_1, median, Q_3, max.)	(55.0, 62.3, 66.2, 70.9, 80.1)	(55.0, 63.0, 67.1, 72.9, 95.9)
Age, $n(\%)$		
[55, 65]	12338 (40.8)	12149 (37.2)
(65, 75]	15409 (50.9)	15003 (46.0)
(75, 95.9]	2498 (8.3)	5477 (16.8)
Prostate-specific antigen (ng/mL)		
(min., Q_1, median, Q_3, max.)	(0.0, 0.7, 1.2, 2.4, 1137.5)	(0.0, 0.7, 1.3, 2.4, 790.9)
Prostate-specific antigen, $n(\%)$		
[0, 4]	26493 (87.6)	29485 (90.4)
(4, 10]	3257 (10.8)	2999 (9.2)
(10, 1137.5]	495 (1.6)	145 (0.4)
Digital rectal exam, $n(\%)$		
Abnormal	3046 (10.1)	1301 (4.0)
Normal	27199 (89.9)	31328 (96.0)
African ancestry, $n(\%)$		
Yes	1286 (4.3)	4464 (13.7)
No	28959 (95.7)	28165 (86.3)
First-degree family history, $n(\%)$		
Yes	2254 (7.5)	5482 (16.8)
No	27991 (92.5)	27147 (83.2)
Prior negative biopsy, $n(\%)$		
Yes	1437 (4.8)	3030 (9.3)
No	28808 (95.2)	29599 (90.7)
Verified, $n(\%)$		
Yes	3813 (12.6)	4185 (12.8)
No	26432 (87.4)	28444 (87.2)

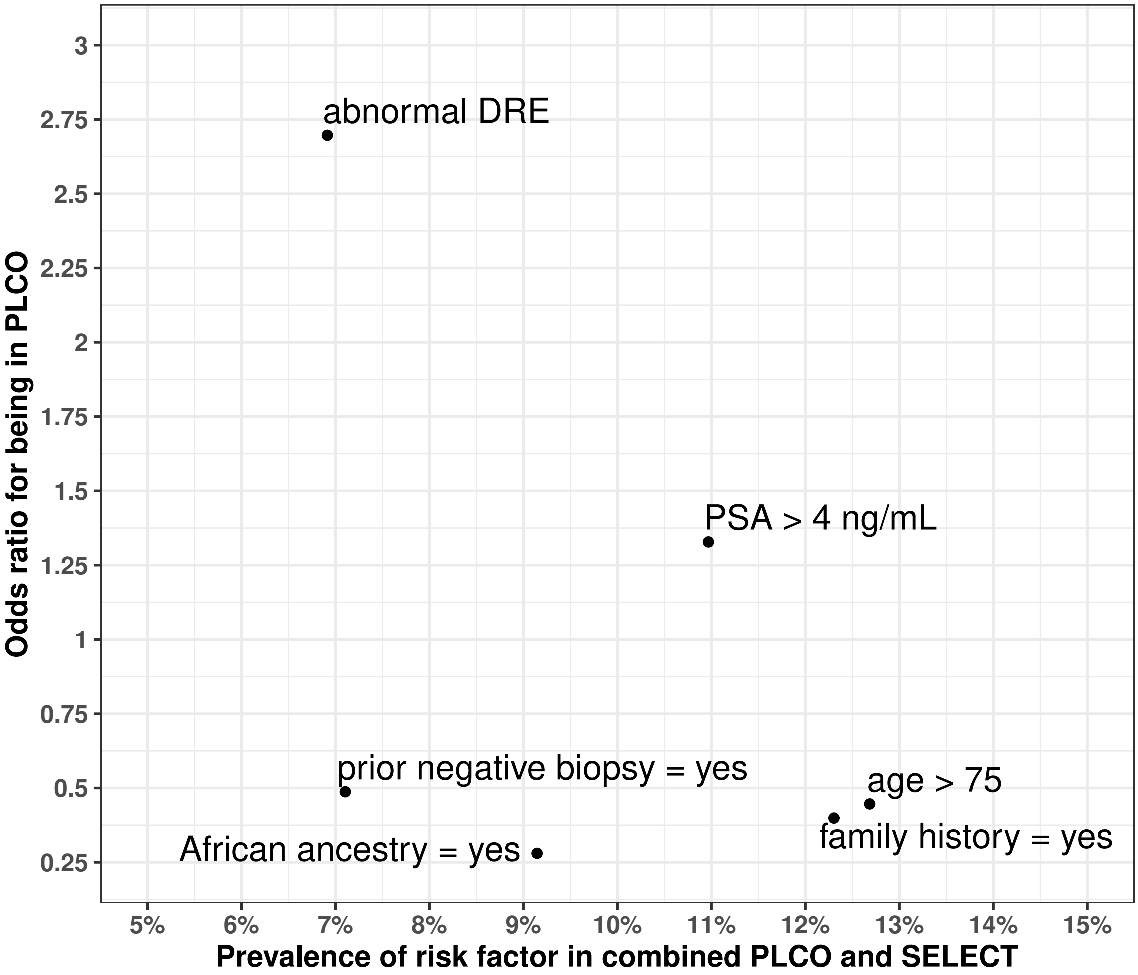

Figure 2.1: Univariable odds ratios for being in PLCO versus prevalence of dichotomized prostate cancer risk factors in 62874 combined participants of PLCO and SELECT (30245 (48.1%) in PLCO). PSA = prostate-specific antigen; DRE = digital rectal exam.

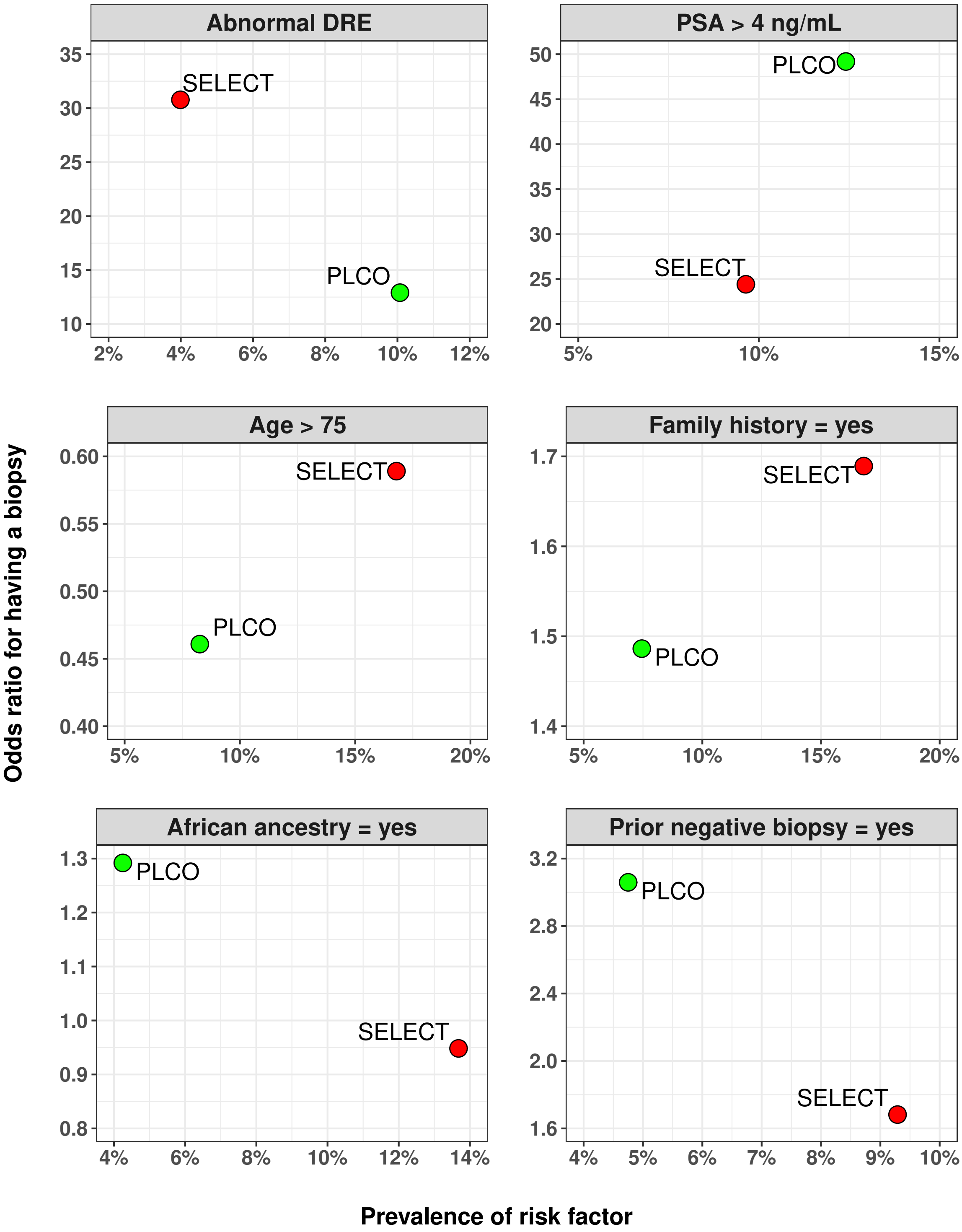

Figure 2.2: Univariable odds ratios for having a biopsy versus the prevalence of dichotomized prostate cancer risk factors evaluating the association between the binary risk factor and the outcome of having a biopsy performed in 30245 PLCO (3813 (12.6%) verified) and 32629 SELECT (4185 (12.8%) verified) participants. PSA = prostate-specific antigen; DRE = digital rectal exam.

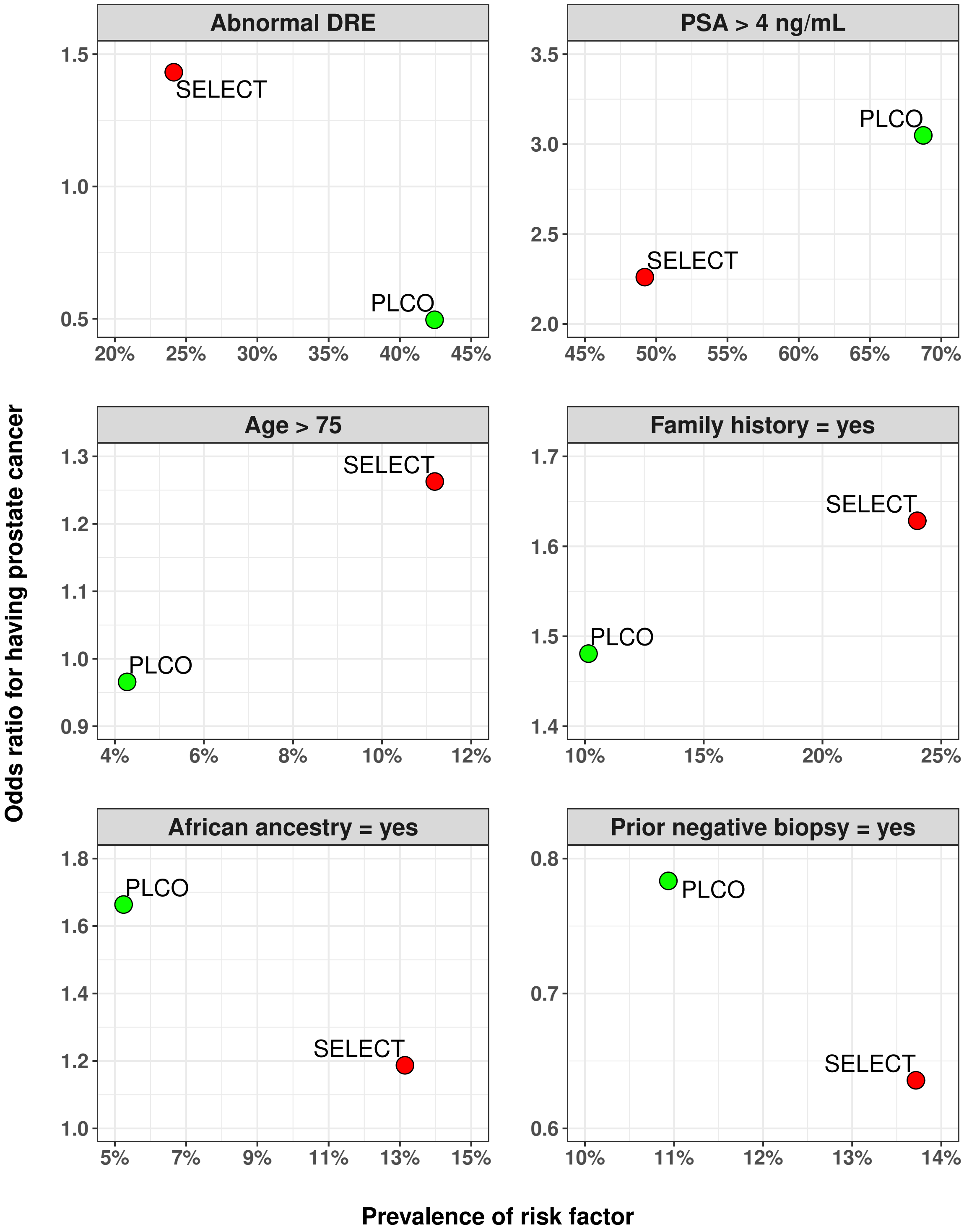

Figure 2.3: Univariable odds ratios for having prostate cancer versus the prevalence of dichotomized prostate cancer risk factors in 3813 PLCO (1833 (48.4%) with cancer) and 4185 SELECT (2028 (48.5%) with cancer) verified participants. PSA = prostate-specific antigen; DRE = digital rectal exam.

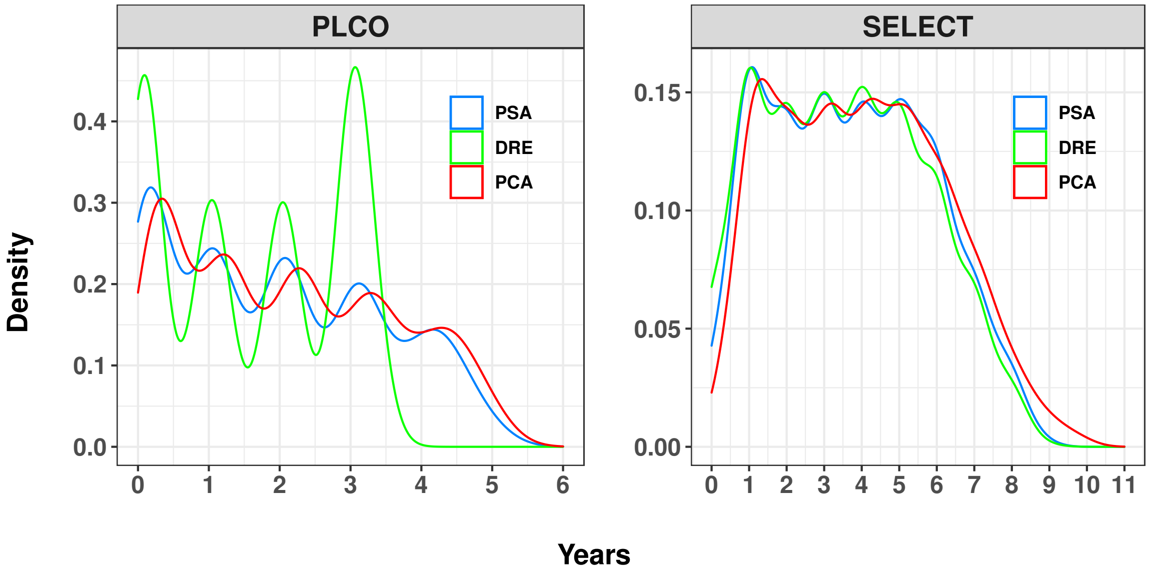

Figure 2.4: Density of the time since study registration until prostate-specific antigen (PSA) test, digital rectal exam (DRE) test, or detection of prostate cancer (PCA) in 1833 PLCO and 2028 SELECT participants with prostate cancer. In SELECT, the time of DRE records taken before the registration date in SELECT is set to be at the registration date.

Scenario 2

For the analysis of the prostate cancer risk with survival endpoint, we define the age at which the first PSA value is measured as the baseline since not all men have a PSA value recorded right at study entry. Men with a prior diagnosis of any cancer or missing values of any risk factors, i.e. PSA, DRE, family history, prior negative biopsy, and African ancestry, are excluded from both PLCO and SELECT. We exclude SELECT men $<$ 55 years and $\geq$ 74 years at baseline since older men are not typically screened for prostate cancer as indicated by the exclusion criterion from PLCO. We further exclude those with PSA $>$ 10 ng/mL at baseline from PLCO. We obtain 29699 PLCO men and 26422 SELECT men for further analysis in the end. In the analysis for time to prostate cancer, we use 5 years after baseline as the censoring time point for both cohorts in this thesis, for which men who have been verified within 5 years after baseline with the biopsy indicator be 1 and 0 otherwise. Similarly, we define the censoring indicator be 1 for men who have prostate cancer within 5 years and 0 otherwise.

Figure 2.5 describes the odds ratio for being in PLCO with respect to dichotomized baseline prostate cancer risk factors in the combined PLCO and SELECT set, where the DRE has been excluded as all participants in SELECT have normal baseline DRE results. Men with first PSA $>$ 2 ng/mL or at an age greater than 65 years at baseline are more likely to be in the PLCO compared to those with lower PSA or at a younger age at baseline. PLCO has a larger percentage of participants with baseline PSA $>$ 2 ng/mL compared to SELECT as shown in Table 2.2, which explains the high odds ratio of being in PLCO for men with

PSA $>$ 2 ng/mL (odds ratio: 1.3). The percentage of men older than 65 years at baseline in PLCO is only slightly higher than that in SELECT and hence, the odds ratio of being in PLCO for age is just moderately higher than one (odds ratio: 1.1). Having prior negative biopsy, family history of cancer, and African ancestry are more likely to be in the SELECT cohort (odds ratios: 0.5, 0.4, and 0.3), which conclusion agrees with the numbers shown in Table 2.2 that the percentages of men with prior negative biopsy, family history, and African ancestry in SELECT are higher than those in PLCO.

From Figure 2.6, participants with a first PSA greater than 2 ng/mL, at an age greater than 65 years, having a family history, with African ancestry, and having prior negative biopsy are more likely to be biopsied within 5 years after baseline, i.e. with odds ratios greater than one, compared to those without for both studies. In PLCO, the odds ratios of having a biopsy for men with abnormal DRE and first PSA $>$ 2 ng/mL are far greater than one (odds ratios: 4.5 and 12.6), which can attribute to the biopsy scheme recommending biopsy men with PSA greater than 4 ng/mL and/or with abnormal DRE. SELECT has the same biopsy scheme and hence, the odds ratio for men with first PSA $>$ 2 ng/mL is around 7.2 in SELECT. The odds ratio for men $>$ 65 years in PLCO is higher than that in SELECT (odds ratios: 1.5 versus 1.2), while the odds ratio for men with family history in SELECT is higher (odds ratios: 1.4 versus 1.6). The odds ratio for participants with African ancestry is slightly higher than that in SELECT. African Americans in PLCO would be biopsied 1.3 times more often than non-African Americans, while around 1.2 times in SELECT. Men in PLCO with prior negative biopsy would be biopsied 2.8 times more often than those without, while around 1.8 times in SELECT. For all risk factors that apply to both cohorts, the odds ratios from both PLCO and SELECT for each risk factor are either above one for both cohorts or below. PLCO referred participants with any risk factors to biopsy more than SELECT, except for family history, which may due to that PLCO cohort was less healthier than SELECT since SELECT required enrolled participants with PSA $<$ 4 and normal DRE. SELECT biopsied more men with family history of cancer than PLCO, possibly due to that SELECT started around 10 years later than PLCO and was more nervous towards occurrence of family history of cancer because public had more knowledge about cancer prevention and knowing that having family history of cancer would be a risky situation to develop cancer.

As for the odds ratio of having prostate cancer within five years after baseline in terms of risk factors, we can see from Figure 2.7 that having a first PSA greater than 2 ng/mL, older than 65 years, having a family history, with African ancestry, and having prior negative biopsy all relate to a higher risk of developing prostate cancer, compared to the others. The odds ratios of having prostate cancer for men in PLCO with PSA $>$ 2 ng/mL, older than 65 years, with African ancestry, and having prior negative biopsy are higher than the respective ones for SELECT, while the odds ratio in PLCO is lower than that in SELECT for men with family history (odds ratios: 1.9 versus 1.7). Men in PLCO with higher PSA are around 20.0 times more often having prostate cancer compared to those with lower PSA, while 8.2 times in SELECT. The large difference in the odds ratios between the two studies may attribute to that PLCO did not exclude participants with PSA $>$ 4 ng/mL at enrollment so that men in SELECT were, in general, healthier than those in PLCO. Specifically, in PLCO, men with abnormal baseline DRE have the odds ratio of having cancer around 2.9. The odds ratios for men with prior negative biopsy differ between the two studies with 1.1 in SELECT and 2.0 in PLCO. The odds ratios for participants with African ancestry are quite similar between SELECT and PLCO (odds ratios: 1.5 versus 1.6), while the odds ratios for the elder people differ between studies (odds ratios: 1.3 versus 1.7).

From Table 2.2, there are no participants with baseline PSA greater than 4 ng/mL or abnormal DRE in SELECT. The proportions of participants with African ancestry, first-degree family history, or prior negative biopsy on PLCO are around half of the corresponding proportions from SELECT, whereas the proportions of being verified and having prostate cancer within 5 years are similar between the cohort as shown in Figure 2.8, i.e. 12.9% in PLCO versus 10.2% in SELECT for being verified and 5.3% in PLCO versus 4.1% in SELECT for having prostate cancer. SELECT had higher percentages of African American was due to the special recruitment grant. More SELECT participants with family history and prio negative biopsy than PLCO, which may due to that SELECT started in 2001 with the population at that time having more knowledge about cancer health condition and care more about cancer prevention than PLCO started in 1993. The log-minus-log transformation is used to construct the point-wise confidence intervals for cumulative incidences to avoid the endpoints of the asymptotic confidence intervals being out of the unit interval [Hosmer Jr et al., 2000]. The incidence of prostate cancer in both cohorts follows much as the patterns of the biopsy. Because PLCO has verified a larger amount of participants, more prostate cancer men in the first year after baseline have been found in PLCO than in SELECT. Overall, PLCO has more prostate cancer cases than SELECT as we expect since the restricted enrollment criteria of PSA $<$ 4 ng/mL and normal DRE prevents participants with poor health conditions from participating in SELECT at the beginning.

Table 2.2: Baseline characteristics verification status within 5 years of 29699 PLCO and 26422 SELECT participants. All p-values from Wilcoxon (Chi-square) tests for numerical (categorical) risk factors and verification status are significant and less than 0.001 except for digital rectal exam as no test was performed on it. Q_1 = the first quartile, Q_3 = the third quartile.

	PLCO (n = 29699)	SELECT (n = 26422)
Age (year)		
(min., Q_1, median, Q_3, max.)	(55.0, 58.1, 62.1, 66.1, 74.0)	(55.0, 58.3, 62.0, 66.4, 74.0)
Age, $n(\%)$		
[55, 60]	10021 (33.7)	9938 (37.6)
(60, 65]	9584 (32.3)	8006 (30.3)
(65, 70]	6827 (23.0)	5767 (21.8)
(70, 74]	3267 (11.0)	2711 (10.3)
Prostate-specific antigen (ng/mL)		
(min., Q_1, median, Q_3, max.)	(0.0, 0.7, 1.1, 2.0, 10.0)	(0.0, 0.7, 1.1, 1.8, 4.0)
Prostate-specific antigen, $n(\%)$		
[0, 1]	13454 (45.3)	12831 (48.6)
(1, 2]	8959 (30.2)	8165 (30.9)
(2, 3]	3465 (11.7)	3543 (13.4)
(3, 4]	1814 (6.1)	1883 (7.1)
(4, 10]	2007 (6.8)	0 (0.0)
Digital rectal exam, $n(\%)$		
Abnormal	2119 (7.1)	0 (0.0)
Normal	27580 (92.9)	26422 (100.0)
African ancestry, $n(\%)$		
Yes	1155 (3.9)	2779 (10.5)
No	28544 (96.1)	23643 (89.5)
First-degree family history, $n(\%)$		
Yes	2288 (7.7)	4623 (17.5)
No	27411 (92.3)	21799 (82.5)
Prior negative biopsy, $n(\%)$		
Yes	1359 (4.6)	2381 (9.0)
No	28340 (95.4)	24041 (91.0)
Verified within 5 years		
Yes	3844 (12.9)	2691 (10.2)
No	25855 (87.1)	23731 (89.8)

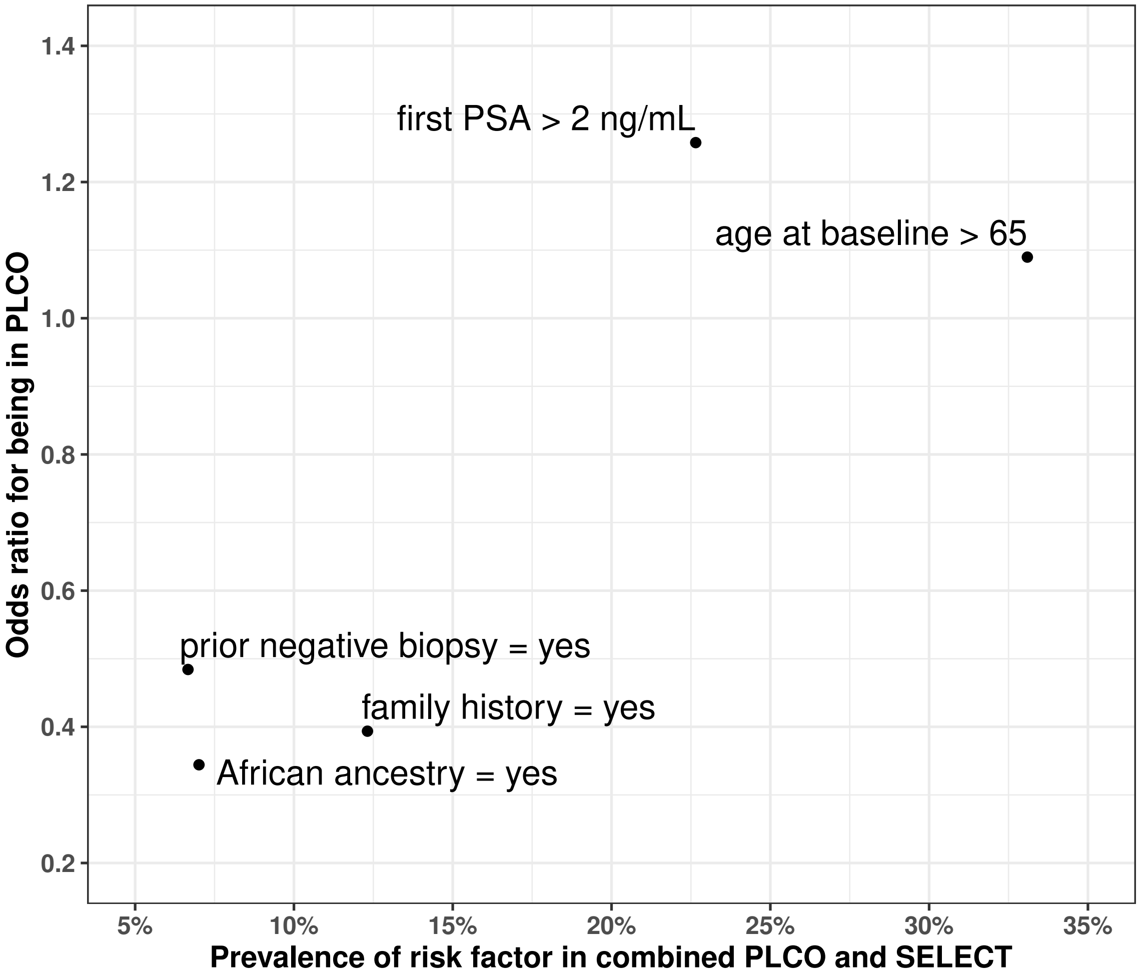

Figure 2.5: Univariable odds ratios for being in PLCO versus prevalence of dichotomized prostate cancer risk factors in 56121 combined participants of PLCO and SELECT (29699 (52.9%) in PLCO). PSA = prostate-specific antigen.

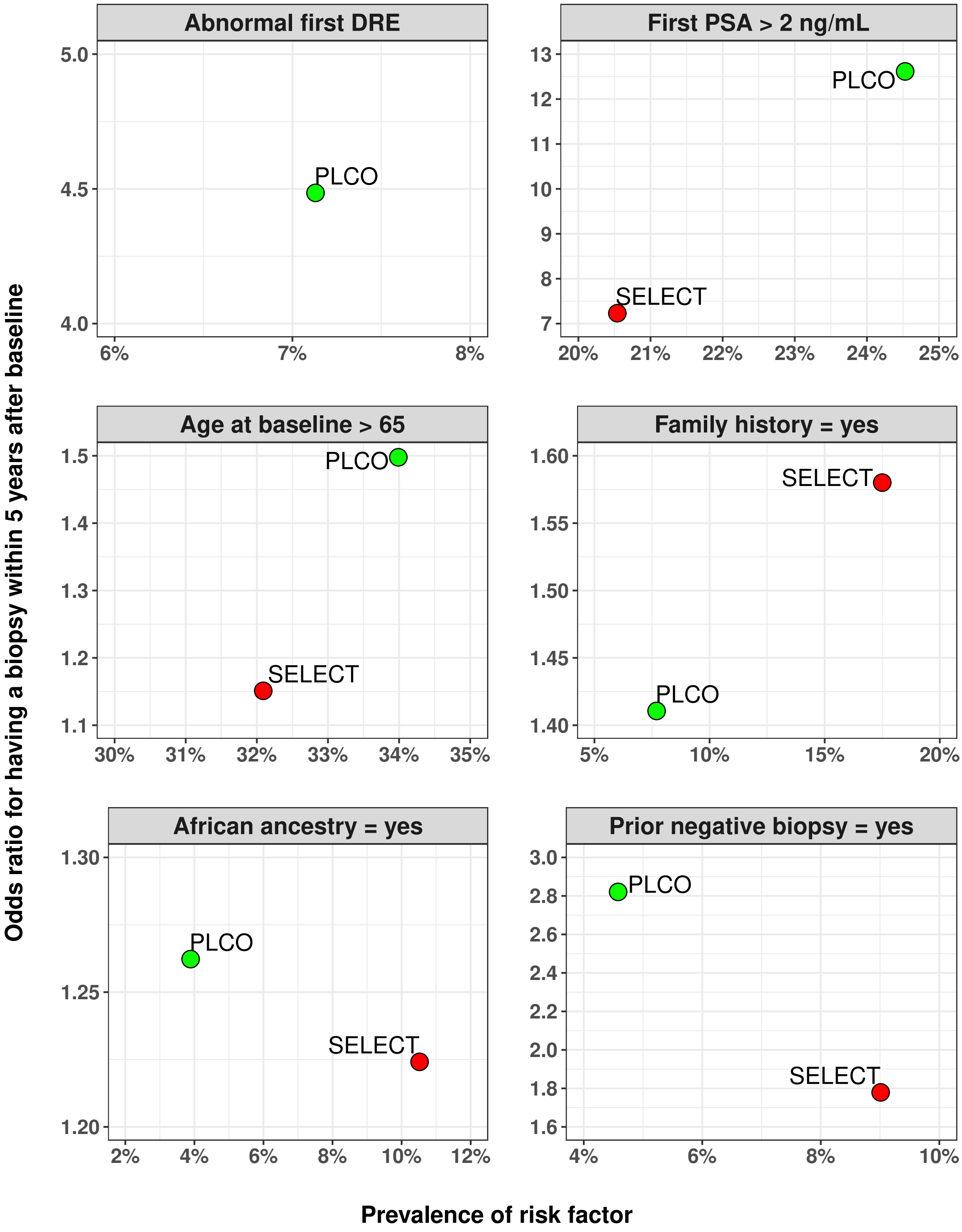

Figure 2.6: Univariable odds ratios for having a biopsy within 5 years after baseline versus the prevalence of dichotomized prostate cancer risk factors evaluating the association between the binary risk factor and the outcome of having a biopsy in 29699 PLCO (3844 (12.9%) verified) and 26422 SELECT (2691 (10.2%) verified) participants. PSA = prostate-specific antigen; DRE = digital rectal exam.

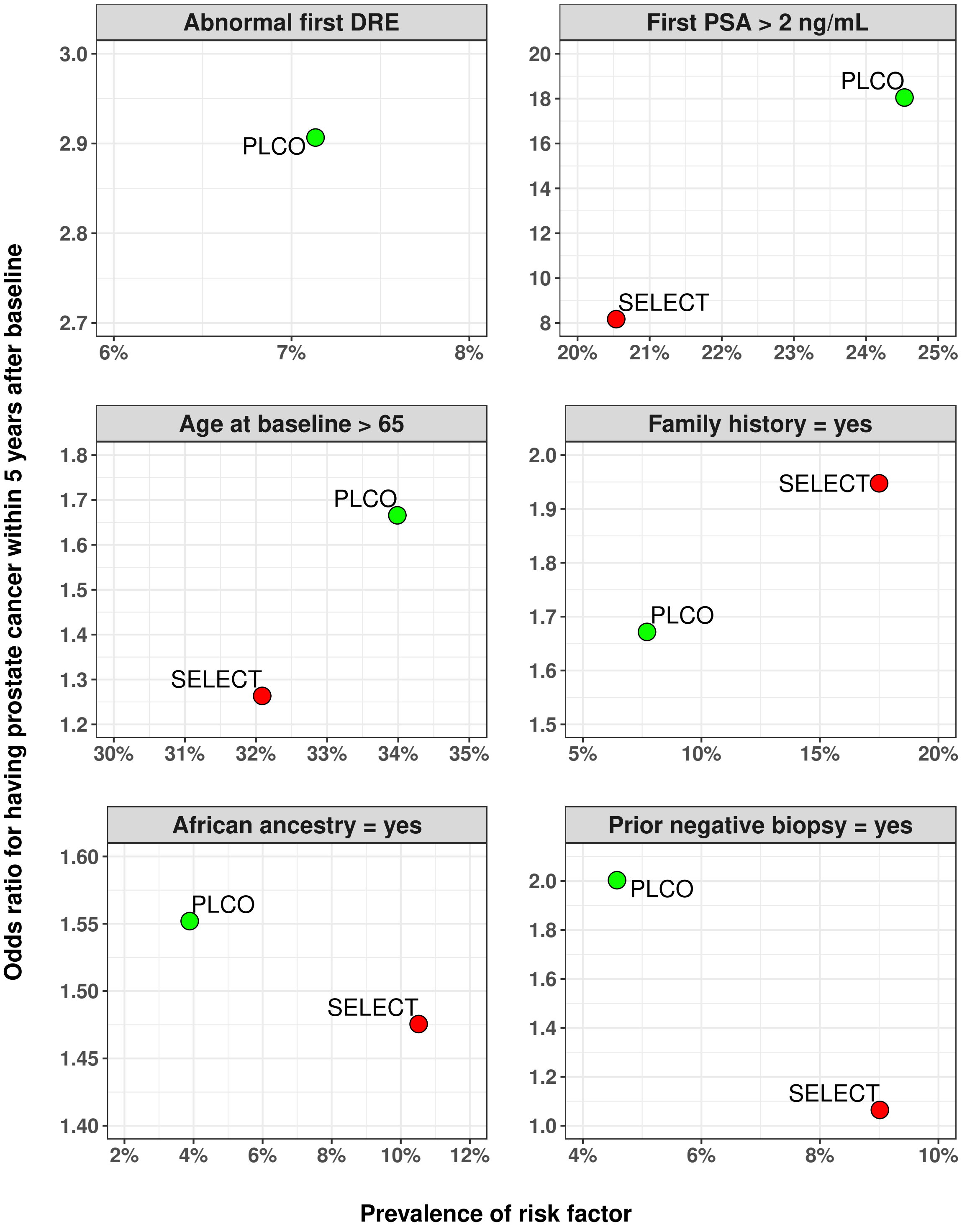

Figure 2.7: Univariable odds ratios for having prostate cancer within 5 years after baseline versus the prevalence of dichotomized prostate cancer risk factors evaluating the association between the binary risk factor and the outcome of having cancer in 29699 PLCO (1564 (5.3%) with cancer) and 26422 SELECT (1093 (4.1%) with cancer) participants. The binary censoring status with value one if a participant developed prostate cancer within 5 years after baseline and zero otherwise is used for calculating the odds ratios. PSA = prostate-specific antigen; DRE = digital rectal exam.

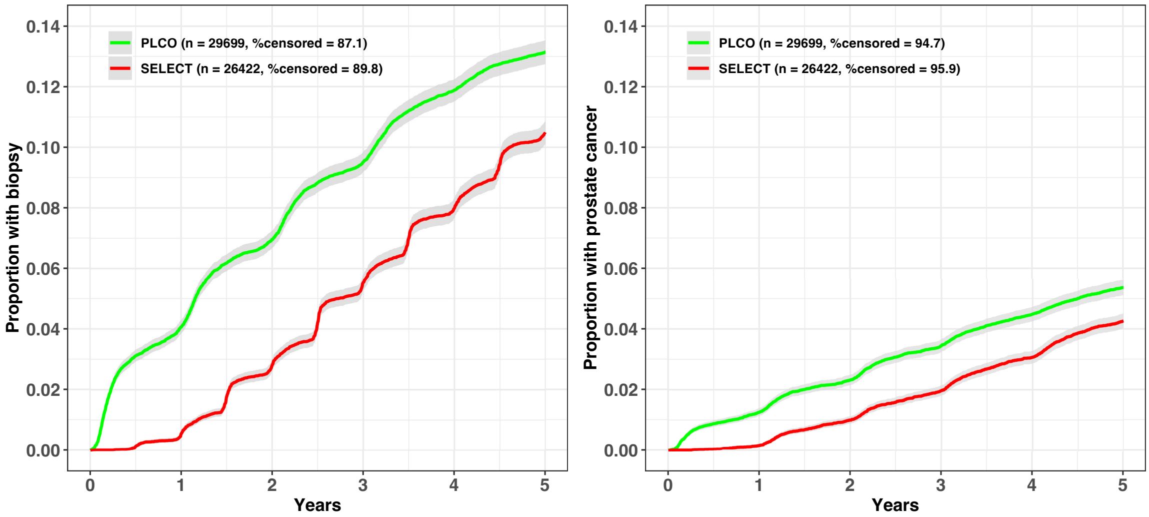

Figure 2.8: Cumulative incidence curves with corresponding log-minus-log confidence intervals for time to first biopsy (left panel) and prostate cancer diagnosis (right panel) using time since first prostate-specific antigen (PSA) measurement as the baseline among 29699 PLCO and 26422 SELECT participants.

3 Standard validation of risk prediction models

In this chapter, we review standard exploratory and quantitative techniques for external validation when individual participant data are available on the training and test sets and illustrate the method with 30245 PLCO and 32629 SELECT participants described in Section 2.3.

3.1 Research context

Once a clinical risk prediction model has been built, we should validate it with external samples for its validity upon samples other than the training population. As discussed previously, the heterogeneity between training and validation cohorts confounds the performance measures in the external validation of risk prediction models, where the heterogeneity could raise from the difference in the distribution of risk factors or variation in the true effects of risk factors upon disease status. Researchers have proposed several methods to assess the extent of heterogeneity between cohorts, such as the AUC from the membership model with a high value reflecting severer heterogeneity in distributions of both risk factors and disease outcome [Debray et al., 2015], or adjusted maximum mean discrepancy metric measuring between-cohort variation in the distributions of risk factors between cohorts [Song et al., 2020]. When the data from both training and validation cohorts are available, we can weight the participants in the validation cohort to resemble the target population or match them based on the propensity score [Ackerman et al., 2019; Powers et al., 2019; Austin, 2011a,b; Austin and Stuart, 2017]. When the training data are not available, which is often the case in the external validation of online risk prediction tools, benchmark values of the external validation metrics serve as the supplement to the usual metric that reveals the impact of variation in risk factors distributions between cohorts in the external validation of the risk prediction models, such as the model-based $c-$index replacing the comparison of observed disease outcomes with the comparison of linear predictors coming from the risk prediction model [Vergouwe et al., 2010; van Klaveren et al., 2016].

One type of bias affecting the external validation of a risk prediction model comes from the fact that we calculate the validation metrics using the data of participants who have been verified by cancer confirmation tests and hence, with known cancer outcomes, but ignore the rest with missing cancer outcomes. The verified participants in the validation cohort may not be randomly sampled from the validation population. Moreover, as discussed before in Section 1.3, the data used to construct a risk prediction model is not a random sample but rather a group of participants with suspicious disease-related symptoms and later, is verified by the gold standard. Using only the data from verified participants whose characteristics are substantially different from the unverified ones causes the so-called verification bias.

When the data from the training cohort are not available, several studies have considered the problem of adjusting the verification bias within the validation cohort in the assessment for the discriminating ability of models with AUC, under either a missing at random (MAR) assumption or with extensions to missing not at random (MNAR) models under potential violations of MAR, where the verification bias occurs when only a part of the participants in a cohort have been verified with gold standard and hence, the disease outcome is missing for unverified ones [Alonzo and Pepe, 2005; Fluss et al., 2009; Buzoianu and Kadane, 2008; Zhang et al., 2018; Zhou and Castelluccio, 2004; Kosinski and Barnhart, 2003]. These methods have developed a disease risk prediction model from verified cases within the validation cohort and used this to impute the disease probability for non-verified participants in the validation cohort. In a review in 2019, 48 publications published between 2005 to 2019 have been found considering imputing the missing outcomes to adjust the verification bias [Umemneku Chikere et al., 2019]. For example, Alonzo and Pepe [2005] imputed the outcomes of the unverified participants with their estimated disease risks from a logistic regression to correct the verification bias under the MAR assumption. To impute the disease risk, other than regression models, we can also apply non-parametric methods to estimate the disease risks that are not subject to misspecification of the models, such as nearest neighbor [Alonzo and Pepe, 2005; He et al., 2009; Adimari and Chiogna, 2015, 2017]. Under the MAR assumption, weighting the verified participants by the inverse of the probability of being verified or combining imputation and weighting utilizing data of all individuals are other approaches to remedy the verification bias [Alonzo and Pepe, 2005; He et al., 2009; He and McDermott, 2012].

When the verification status is related to the unobserved data, MAR does not hold. In this case, some verification bias adjustment approaches incorporate the association between verification status and observed disease outcome in the verification model, such as the doubly robust estimator for AUC [Rotnitzky et al., 2006; Fluss et al., 2009; Zhang et al., 2018] and likelihood approach [Liu and Zhou, 2010; Zhou and Castelluccio, 2004]. They all apply a pre-specified parameter to quantify the extent of association between verification status and the disease outcomes, such as the log odds ratios of having the disease for verified versus unverified ones under the same risk factors and diagnostic test result levels. Other eligible verification bias adjustment methods under MNAR include propensity score adjustment using the instrumental variable [Yu et al., 2018] and Bayesian approaches modeling verification probability with disease outcome [Buzoianu and Kadane, 2008].

Because the adjustment of verification bias is mainly investigated in the attempt to perfectly evaluate the accuracy of the diagnostic test, i.e. comparing the results from the test versus the true outcomes verified by the gold standard to determine if this test can distinguish the case versus non-case correctly, all publications focus only on addressing the bias in the estimation of AUC or sensitivity and specificity showing the discriminating ability of the diagnostic test as far as we can see.

Here, we utilize the individual participant data from unverified participants in the validation cohorts in addition to the verified ones to adjust the verification bias in the external validation results of risk models under the MAR assumption. Both verified and unverified participants should have risk factors available, whereas clinical outcomes are only available for verified participants. We illustrate the process with data from PLCO and SELECT, where the PLCO is used as the training cohort and SELECT as the external validation cohort.

3.2 Notations and metrics

We outline the usual model training and validation framework in Figure 3.1. Rather than considering the training and validation cohorts separately, we envision them as arising from a pool of individuals, all of which have the risk predictors X for a disease measured. In other words, the training and validation cohorts should share the same risk factors. We let T denote which cohort the individuals have been selected into, with $T = 1$ the training cohort for a prediction model and $T = 0$ the cohort to validate it. Selection into a cohort typically depends on risk factors X, often specified as eligibility criteria, which is indicated in Figure 3.1. Once in a cohort, whether or not the individual is verified for the disease, V, also depends on the risk factors X typically. Standards for referral for verification often vary across cohorts and thus V may depend on X and T. Finally, we assume that the disease status D is inherent to the individual, depending only on their risk factors X, and not on the selection T or verification V mechanisms. Note that if $V = 0$ then D is unobserved.

We assume a risk prediction model is built relating the disease outcomes to the risk factors in the training cohort either by just using the verified individuals who have outcomes available, as in the case of the prostate cancer application, or by including unverified individuals additionally via multiple imputation for missing outcome data. The modeling yields the coefficients for a risk function $R(X)$ that can then be applied to the risk factors for individuals in the external validation cohort. For simplicity, we assume that all individuals in both cohorts have the same risk factors X measured, though missing risk factors could be filled in by some imputation procedures.

In the external validation of the built model, the AUC and the CIL are two common metrics used for evaluating the discrimination and calibration respectively. Using the notation in Figure 3.1, the typical CIL summarizing the expected discrepancy between predicted risk

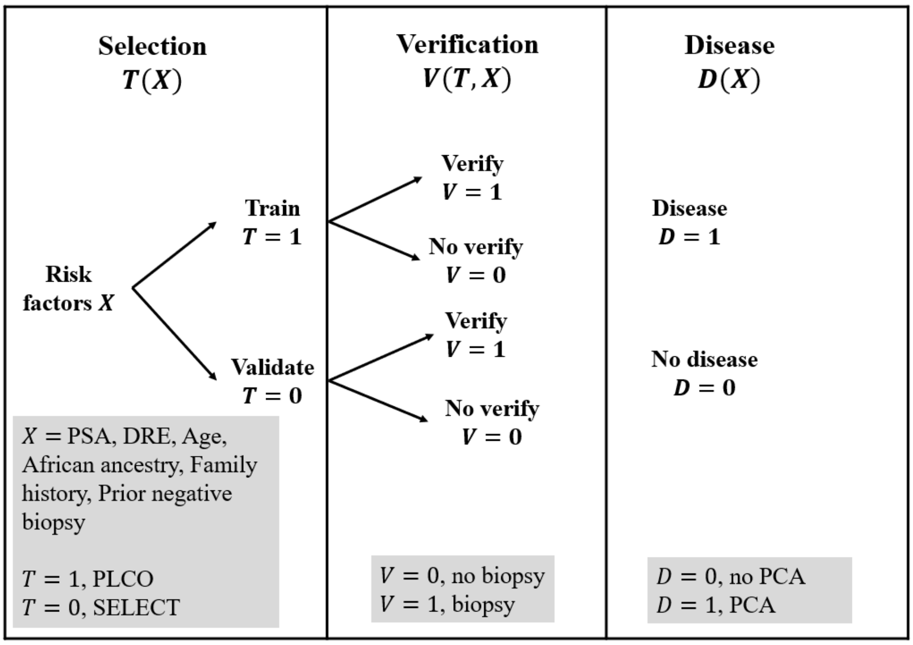

Figure 3.1: Data collection processes of a validation and training set. Lower panels indicate definitions for the prostate cancer (PCA) application with PSA: prostate-specific antigen, DRE: digital rectal exam.

and disease status in the validation set ($T = 0$) is

$$CIL = E(R(X) - D|T = 0, V = 1). \tag{3.1}$$

Assuming that (X_i, T_i, V_i, D_i) are independent and identical distributed with distribution $F_{X,T,V,D}$ for $i = 1, \ldots, N = N_0 + N_1$ individuals across both the training (N_1) and validation (N_0) cohorts, we can approximate the CIL by the sample average as

$$\widehat{CIL} = \frac{\sum_{i=1}^{N_0}(R(X_i) - D_i)V_i}{\sum_{i=1}^{N_0} V_i}. \tag{3.2}$$

which is a consistent estimator of CIL by the law of large number when the effective sample size $\sum_{i=1}^{N_0} V_i \to \infty$. Specifically, let $I(A)$ be the indicator function with value one when event A occurs and zero otherwise, we re-write CIL and the corresponding estimator as

$$\begin{aligned} CIL &= E(R(X)|VI(T = 0) = 1) - E(D|VI(T = 0) = 1), \\ \widehat{CIL} &= \frac{\sum_{i=1}^{N_0}(R(X_i) - D_i)V_i}{\sum_{i=1}^{N_0} V_i} = \frac{\sum_{i=1}^{N}(R(X_i) - D_i)V_i I(T_i = 0)}{\sum_{i=1}^{N} V_i I(T_i = 0)}. \end{aligned}$$

Since the risk function R is real-value continuous function of X, we have

$$\frac{1}{N}\sum_{i=1}^{N} R(X_i)V_i I(T_i = 0) \xrightarrow[N\to\infty]{p} E[R(X)VI(T = 0)]; \quad \frac{1}{N}\sum_{i=1}^{N} V_i I(T_i = 0) \xrightarrow[N\to\infty]{p} P(VI(T = 0) = 1)$$

by law of large numbers, where $\xrightarrow{p}$ denotes convergence in probability. Assuming $P(VI(T = 0) = 1) > 0$, by continuous mapping theorem, we have

$$\frac{\frac{1}{N}\sum_{i=1}^{N} R(X_i)V_iI(T_i = 0)}{\frac{1}{N}\sum_{i=1}^{N} V_iI(T_i = 0)} \xrightarrow[N\to\infty]{p} \frac{E[R(X)VT]}{P(VI(T = 0) = 1)} = E[R(X)|VI(T = 0) = 1].$$

The above equality is due to the fact that $E[R(X)VI(T = 0)] = \int R(X)P(X, VI(T = 0) = 1)dX$, where the equality holds by dividing the function inside the integral with $P(VI(T = 0) = 1)$. Here by assuming $P(VI(T = 0) = 1) > 0$ we ensure that the verified participants exist among the population with certain positive probability. When the sample size N increases to infinity, the number of verified individuals among N goes to infinity as well, i.e. $\sum_{i=1}^{N_0} V_i \to \infty$ when $N \to \infty$. Similarly, we can show

$$\frac{\frac{1}{N}\sum_{i=1}^{N} D_iV_iI(T_i = 0)}{\frac{1}{N}\sum_{i=1}^{N} V_iI(T_i = 0)} \xrightarrow[N\to\infty]{p} \frac{E[DVI(T = 0)]}{P(VI(T = 0) = 1)} = E[D|VI(T = 0) = 1].$$

Then, the sum of two consistent estimators is consistent, i.e.

$$\frac{\frac{1}{N}\sum_{i=1}^{N} R(X_i)V_iI(T_i = 0)}{\frac{1}{N}\sum_{i=1}^{N} V_iI(T_i = 0)} - \frac{\frac{1}{N}\sum_{i=1}^{N} D_iV_iI(T_i = 0)}{\frac{1}{N}\sum_{i=1}^{N} V_iI(T_i = 0)} \xrightarrow[N\to\infty]{p} E[R(X)|VI(T = 0) = 1] - E[D|VI(T = 0) = 1]. \quad (3.3)$$

So that $\widehat{CIL}$ is a consistent estimator of CIL.

By the central limit theorem, the variance of $\widehat{CIL}$ is $\sigma^2/\sum_{i=1}^{N_0} V_i$, where $\sigma^2 = Var(R(X) - D|T = 0, V = 1)$, and its distribution is asymptotically normal. The variance σ^2 can be estimated by the sample variance of $R(X) - D$ among verified participants in the validation set, yield the estimate variance of $\widehat{CIL}$ as

$$\widehat{var}(\widehat{CIL}) = \frac{1}{\sum_{i=1}^{N_0} V_i}\left(\frac{\sum_{i=1}^{N_0}(R(X_i) - D_i)^2V_i}{\sum_{i=1}^{N_0} V_i} - \widehat{CIL}^2\right). \quad (3.4)$$

Discrimination begins with true positive rates (TPRs) and false positive rates (FPRs) for rules that would test positive for disease when $R > c$. They can be computed for all possible thresholds $c \in [0, 1]$, and are commonly estimated among the verified participants in the validation set by:

$$TPR(c) = P(R(X) > c|V = 1, T = 0, D = 1),\ FPR(c) = P(R(X) > c|V = 1, T = 0, D = 0),$$
$$\widehat{TPR}(c) = \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)V_iD_i}{\sum_{i=1}^{N_0} V_iD_i},\ \widehat{FPR}(c) = \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)V_i(1 - D_i)}{\sum_{i=1}^{N_0} V_i(1 - D_i)}, \quad (3.5)$$

where $I(\cdot)$ is an indicator function equal to 1 if the argument inside the bracket holds and 0 otherwise. Higher values of the TPR and lower values of the FPR are desirable, though these have a trade-off that depends on c. To summarize overall c, the receiver-operating-characteristic (ROC) curve plots $FPR(c)$ on the x-axis versus $TPR(c)$ on the y-axis for all $c \in [0,1]$. Higher ROC curves that have maximal TPRs close to 1 for all FPRs indicate better discrimination. The AUC summarizes the ROC curve as the area underneath it, with higher values close to 1 indicating better discrimination. We can calculate the corresponding AUC by trapezoidal rule and obtain its standard error via bootstrapping.

The estimators in (3.5) are consistent estimators for the corresponding rates. Applying the same notation as before in calibration, we take TPR as an example and FPR follows similarly. By law of large numbers, we have

$$\frac{1}{N}\sum_{i=1}^{N} I(R(X_i) > c)V_i I(T_i = 0)D_i \xrightarrow[N\to\infty]{p} E[I(R(X) > c)VI(T = 0)D]$$
$$\frac{1}{N}\sum_{i=1}^{N} V_i I(T = 0)D_i \xrightarrow[N\to\infty]{p} P(VI(T = 0)D = 1).$$

Then, for $\widehat{TPR}(c)$, we can show the consistency of the estimator as

$$\begin{aligned}\frac{\frac{1}{N}\sum_{i=1}^{N} I(R(X_i) > c)D_i V_i I(T_i = 0)}{\frac{1}{N}\sum_{i=1}^{N} V_i I(T_i = 0)D_i} \xrightarrow[N\to\infty]{p} & \quad \frac{E[I(R(X) > c)VI(T = 0)D]}{P(VI(T = 0)D = 1)} \\ = & \quad P(I(R(X) > c)|V = 1, T = 0, D = 1)\end{aligned}$$

by continuous mapping theorem given $P(VI(T = 0)D = 1) > 0$, i.e. the prevalence of disease ($D = 1$) among the verified participants ($V = 1$) in the validation cohort ($T = 0$) is greater than 0. According to central limit theorem, $\widehat{TPR}(c)$ distributes asymptotically to a normal distribution with variance of

$$\frac{1}{\sum_{i=1}^{N_0} V_i D_i}\left[\frac{\sum_{i=1}^{N_0} I(R(X_i) > c)V_i D_i}{\sum_{i=1}^{N_0} V_i D_i}\left(1 - \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)V_i D_i}{\sum_{i=1}^{N_0} V_i D_i}\right)\right], \tag{3.6}$$

which is the sample variance of $I(R(X) > c)$ among verified participants with the disease in the validation cohort divided by the number of verified participants with disease among N_0. Here, the $I(R(X) > c)$ follows a Bernoulli distribution with an estimated success probability

$$P(I(R(X) > c) = 1|V = 1, T = 0, D = 1) = \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)V_i D_i}{\sum_{i=1}^{N_0} V_i D_i}. \tag{3.7}$$

Missing-at-random adjustments

Following previous approaches, we can extend the validation to include the unverified participants in the validation set under the assumption their missing outcomes follow a missing-at-random (MAR) mechanism [Alonzo and Pepe, 2005]. The method proceeds by substituting the missing cancer status D_i for unverified individuals in the validation set with an estimated disease probability $p_i = P(D = 1|X_i, T = 0)$ from a model for disease status built on the verified participants in the validation set. The adjustment of $\widehat{CIL}$ for MAR becomes

$$\widehat{CIL}_{MAR} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left[(R(X_i) - D_i)V_i + (R(X_i) - \hat{p}_i)(1 - V_i) \right]. \tag{3.8}$$

All other metrics can be similarly adjusted, such as the unweighted and weighted TPRs and FPRs:

$$\widehat{TPR}_{MAR}(c) = \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)[V_i D_i + (1 - V_i)\hat{p}_i]}{\sum_{i=1}^{N_0} V_i D_i + (1 - V_i)\hat{p}_i},$$

$$\widehat{FPR}_{MAR}(c) = \frac{\sum_{i=1}^{N_0} I(R(X_i) > c)[V_i(1 - D_i) + (1 - V_i)(1 - \hat{p}_i)]}{\sum_{i=1}^{N_0} V_i(1 - D_i) + (1 - V_i)(1 - \hat{p}_i)},$$

and the AUCs under MAR assumption follow. We can construct the confidence intervals using the bootstrapping as well.

Weighted estimators

The above estimators can be weighted to account for training and validation cohort differences as will be seen in Chapter 4. We show here the large sample properties of the weighted estimates. Specifically, among n validation participants, the estimators are in the form of

$$\frac{\hat{S}_n}{\hat{Z}_n} = \frac{1/n \sum_{i=1}^{n} g(X_i, D_i, V_i) w(X_i, \hat{\beta})}{1/n \sum_{i=1}^{n} h(X_i, D_i, V_i) w(X_i, \hat{\beta})} \tag{3.9}$$

for certain bounded functions g, h for X, D, V and weights function w of X. The large sample properties of estimators depend on the behaviour of both the variables and estimated coefficients ($\hat{\beta}$) for X in the model for the weights ($w(X_i, \hat{\beta})$) and the estimated coefficients for the probability of having cancer in the validation cohort $P(D = 1|T = 0, X)$ when impute the missing outcome for the unverified participants under MAR assumption. β is often known as nuisance parameter.

To investigate the limiting behavior of the proposed estimators, we notice that the coefficients (β) are estimated using the observed data first and being embedded to calculate the weighted estimators afterward. We can view such a two-step estimating problem as a one-step problem such that the estimators and the coefficients are computed simultaneously, which process consolidates two steps into one has been justified by Newey [1984]. In this case, to find the estimators for CILs is to solve the summation of estimating equations over n participants in the validation cohort, i.e. solving

$$\sum_{i=1}^{n} \left\{ \left[(R(X_i) - D)V_i + k(R(X_i) - \rho_i)(1 - V_i) \right] w(X_i, \beta) - \pi_1 \right\} = 0 \tag{3.10}$$

$$\sum_{i=1}^{n} \left\{ \left[V_i + k(1 - V_i) \right] w(X_i, \beta) - \pi_2 \right\} = 0 \tag{3.11}$$

for π_1 and π_2, where $\rho_i = P(D_i = 1|X_i, T_i = 0)$ and π_1, π_2 correspond to the estimators for the numerator and denominator of CILs, respectively. When $w(X, \beta)$ be a constant 1 and $k = 0$, π_1/π_2 is the estimator for CIL, while the estimator for CIL_{MAR} when $k = 1$. Later, the weights $w(X, \beta)$ could be a probability depending on X (see Chapter 4). Similarly for TPRs, we solve the following summation of estimating equations:

$$\sum_{i=1}^{n} \{I(R(X_i) > c)\,[V_i D_i + k(1 - V_i)\rho_i]\,w(X_i, \beta) - \pi_3\} = 0 \tag{3.12}$$

$$\sum_{i=1}^{n} \{[V_i D_i + k(1 - V_i)\rho_i]\,w(X_i, \beta) - \pi_4\} = 0 \tag{3.13}$$

for π_3 and π_4, obtaining the estimators for the numerator and denominator of TPRs, respectively. π_3 corresponds to $P(R(X) > c, V = 1, D = 1, T = 0)$, while π_4 corresponds to $P(V = 1, D = 1, T = 0)$. When $k = 0$ or 1 and $w(X, \beta)$ be constantly 1 or a probability depending on X, we have the estimators for TPRs. We can define the estimating equations for the numerator (π_5) and denominator (π_6) for FPRs analogously as (3.12) and (3.13).

Because the estimation for CILs does not relate to the estimation of TPRs and FPRs, we treat the estimations of calibration and discrimination measures as distinct processes and hence, their large sample properties can be proved separately. We now take the CILs as the example showing the limiting properties, while the properties for TPRs and FPRs can be checked similarly. Let $\pi = (\pi_1, \pi_2)^\intercal$ corresponding to the numerator and denominator of CILs and $\theta = (\pi, \beta)^\intercal$. The superscript "$\intercal$" refers to transposition. The coefficients (β) are usually estimated from certain estimating equation as well. We then let $U(\theta)$ be an estimating equation for θ and $U_n(\theta) = \sum_{i=1}^{n} U_i(\theta)$, where $U_i(\theta) = (U_i^{\pi}(\theta), U_i^{\beta}(\theta))^\intercal$. Solving $U_n(\theta) = 0$ for θ gives the estimators as well as the estimated coefficients all at once. We denote $\hat{\theta}_n$ as the solution for $U_n(\theta) = 0$. Then, we just need to show the large sample properties of the estimated $\hat{\theta}_n$. Let $Y_i = (X_i, T_i, V_i, D_i)$ be independent identical distributed (i.i.d.) following distribution F_Y for $i = 1, \dots, n$ and $U_i(\beta)$ is i.i.d. across i as well. T, V, D are dichotomous random variables. With the MAR assumption, we can impute the outcome for the unverified participants with $P(D = 1|T = 0, X)$. We now assume the models for $P(D = 1|T = 0, X)$ and the weights are correctly specified and introduce the following additional assumptions:

H1: $\theta = (\pi, \beta)^\intercal \in \Theta$ a closed and bounded parameter space,

H2: exists a unique $\theta_0 \in \Theta$ such that $E[U(\theta_0)] = 0$,

H3: $U(\theta)$ is differentiable and hence, a continuous function with respect to θ,

H4: the expectation of the absolute estimating equation $|U(\theta)|$ over Y: $E_Y[\sup_{\theta\in\Theta} |U(\theta)|] < \infty$.

H1 is a reasonable assumption since π is bounded and β cannot take infinite value in reality. H2 assumes that there exists a unique solution θ_0 for estimating equation $E[U(\theta)] = 0$.

H3 is satisfied in our case since $U(\theta)$ is continuous in π with $\partial U^{\pi}(\theta)/\partial\pi = -1$. When the weights come from logistic regression based on X with parameter β, $\partial U^{\pi}(\theta)/\partial\beta$ also exists because it relates to the derivative of $w(X,\beta) = e^{\beta X}/(1+e^{\beta X})$ with respective to β. When using maximum likelihood estimation to obtain $\hat{\beta}$ from a logistic regression

$$ln\left(\frac{w(X,\beta)}{1-w(X,\beta)}\right) = \beta X, \tag{3.14}$$

the summation of the estimating equations for β is the derivative of log-likelihood function ($l(\beta)$) from the logistic regression over β, i.e.

$$\frac{\partial l(\beta)}{\partial\beta} = \sum_{i=1}^{n} V_i X_i - w(X_i,\beta)X_i, \tag{3.15}$$

where $w(X_i,\beta) = e^{\beta X_i}/(1+e^{\beta X_i})$ and the estimated $\hat{\beta}$ is the solution when (3.15) is 0. (3.15) is differentiable at β since $\partial w(X_i,\beta)/\partial\beta$ exists, while its derivative with respect to π is 0 because it does not involve π. H4 focuses on the expectation over Y, i.e. over all random variables in the estimating equation $U(\theta)$ other than the nuisance parameter β. It assumes that the expectation of $|U(\theta)|$ with θ varying in the parameter space is finite, i.e. first calculating the expectation of $|U(\theta)|$ over Y fixing θ and then, the maximum of such expectations across different θ values is assumed to be finite. This condition holds in our case since the estimating equation for π has bounded value and whose expectation is finite. As long as the covariates (X) in the logistic regression have finite expectations, the expectation of estimating equation for β is also bounded as we can see from (3.15).

The idea of showing consistency is that we have known that $\hat{\theta}_n$ solves $U_n(\theta) = 0$ and θ_0 solves $E[U(\theta)] = 0$. When $U_n(\theta)$ converges uniformly to $E[U(\theta)]$, we can show $\hat{\theta}_n$ converges to θ_0 in probability when $n \to \infty$ under the above assumptions. With H3 and H4 and the uniform law of large numbers [Newey and McFadden, 1994, Lemma 2.4], we have that $E[U(\theta)]$ is continuous over $\theta \in \Theta$ and

$$\frac{1}{n}\sum_{i=1}^{n} U_i(\theta) \xrightarrow[n\to\infty]{p} E[U(\theta)] \text{ uniformly.} \tag{3.16}$$

With these two conclusions above and H1 - H2, by Theorem 2.1 in Newey and McFadden [1994], we conclude that

$$\hat{\theta}_n \xrightarrow[n\to\infty]{p} \theta_0, \tag{3.17}$$

which means the estimated $\hat{\theta}_n$ from solving $U_n(\theta) = 0$ is a consistent estimator of its true value θ_0. Therefore, the estimated CILs by solving the respective estimating equations are consistent estimators of their corresponding true values because the estimated numerators and denominators are components of $\hat{\theta}_n$ that is a consistent estimator for θ_0, the true parameter vector solving $E[U(\theta)] = 0$.

Towards the limiting distribution, we make the following further assumptions:

H5: $var\{U(\theta_0)\}$ is positive and finite,

H6: $E_Y[\sup_{\theta\in\Theta}|\frac{\partial U(\theta)}{\partial\theta}|] < \infty$,

H7: $E[\frac{\partial U(\theta)}{\partial\theta}]$ is non-singular and invertible.

We expand the estimating equation $U_n(\hat{\theta}_n)$ around the true value θ_0 with Taylor series truncating at the second term according to mean value theorem as

$$\begin{aligned} 0 &= \sqrt{n}U_n(\hat{\theta}_n) && (3.18)\\ &= \sqrt{n}U_n(\theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)\frac{\partial U_n(\tilde{\theta}_n)}{\partial\theta}, && (3.19)\end{aligned}$$

where $\tilde{\theta}_n = \alpha\hat{\theta}_n + (1-\alpha)\hat{\theta}_n$ with $\alpha \in (0,1)$. Since $\hat{\theta}_n \xrightarrow{p} \theta_0$, we have $\tilde{\theta}_n \xrightarrow{p} \theta_0$ as well. By uniform law of large numbers and H6, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i(\theta)}{\partial\theta} \xrightarrow[n\to\infty]{p} E\left[\frac{\partial U(\theta)}{\partial\theta}\right] \text{ uniformly.} \tag{3.20}$$

Then,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i(\tilde{\theta}_n)}{\partial\theta} \xrightarrow[n\to\infty]{d} E\left[\frac{\partial U(\theta_0)}{\partial\theta}\right] \tag{3.21}$$

holds by similar argument as showing consistency above. Given H5 and by central limit theorem, we have

$$\sqrt{n}[\frac{1}{n}U_n(\theta_0)] \xrightarrow[n\to\infty]{d} N(0, var[U(\theta_0)]). \tag{3.22}$$

Due to H7 that $E\left[\frac{\partial U(\theta_0)}{\partial\theta}\right]^{-1}$ exists, we have the asymptotic normality of $\hat{\theta}_n$ by Slutzky theorem as

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) &= \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i(\tilde{\theta}_n)}{\partial\theta}\right]^{-1}\sqrt{n}\left(\frac{1}{n}U_n(\theta_0)\right)\\ &\xrightarrow[n\to\infty]{d} N\left(0, E\left[\frac{\partial U(\theta_0)}{\partial\theta}\right]^{-1} var\{U(\theta_0)\}\left(E\left[\frac{\partial U(\theta_0)}{\partial\theta}\right]^{\intercal}\right)^{-1}\right). \end{aligned} \tag{3.23}$$

Therefore, the ratio constructing with $\pi_j, j = 1, \ldots, 6$ follows a normal distribution asymptotically as well, where the asymptotic variance can be constructed by multivariate delta method similarly as in Alonzo and Pepe [2005]. For example, for the ratio π_1/π_2 corresponding to CILs, the asymptotic variance is

$$\frac{\partial h(\theta)}{\partial\theta}\Sigma\left(\frac{\partial h(\theta)}{\partial\theta}\right)^{\intercal},$$

where $h(\theta) = \pi_1/\pi_2$ and $\partial h(\theta)/\partial\theta$ is a row vector. The Σ is the variance of the distribution in (3.23). The asymptotic variance of the estimator of CILs relates to the variances of π_1, π_2, and β as well as their covariance between each others. If we let $\pi = (\pi_3, \pi_4, \pi_5, \pi_6)$, we can show the properties for the estimators for TPRs and FPRs. Their asymptotic variances follow by replacing the $h(\theta)$ in (3.2) with $(\pi_3/\pi_4, \pi_5/\pi_6)^\intercal$, obtaining for TPRs and FPRs simultaneously.

Now we give an approximation of the asymptotic variance of the estimated ratio. In our case, $h(X, D, V)$ is always 1 and $\hat{Z}_n$ in (3.9) becomes $1/n \sum_{i=1}^n w(X_i, \hat{\beta})$. Since we are lack of exact analytic expression of the asymptotic variance of the ratio estimator whose variance is usually obtained via bootstrapping in practice, one approximation of this variance is [Gatz and Smith, 1995]:

$$\frac{n}{(n-1)\left(\sum_{i=1}^n w(X_i, \hat{\beta})\right)^2} \left\{ \sum_{i=1}^n w^2(X_i, \hat{\beta}) \left(g(X_i, D_i, V_i) - \frac{\hat{S}_n}{\hat{Z}_n} \right)^2 \right\}, \tag{3.24}$$

which Gatz and Smith [1995] claimed to be a reasonable estimation for the variance compared to the one derived via bootstrap. Here, we approximate the variance through Taylor expansion of ratio. Specifically, the variance of Taylor approximation of ratio S/Z around the respective mean values (μ_S, μ_Z) is

$$var(S/Z) \approx \frac{1}{\mu_Z^2} \left[var(S) - 2\frac{\mu_S}{\mu_Z} cov(S, Z) + \frac{\mu_S^2}{\mu_Z^2} var(Z) \right]. \tag{3.25}$$

The variance of S is approximated by the sample variance as

$$\frac{1}{n} \sum_{i=1}^n \widehat{var}(g(X_i, V_i, D_i) w(X_i, \hat{\beta})) = \\ \frac{1}{n(n-1)} \sum_{i=1}^n \left(g(X_i, V_i, D_i) w(X_i, \hat{\beta}) - \frac{\sum_{i=1}^n g(X_i, V_i, D_i) w(X_i, \hat{\beta})}{n} \right)^2.$$

The covariance between S and Z is approximated by

$$\frac{1}{n} \sum_{i=1}^n \widehat{cov}(g(X_i, V_i, D_i) w(X_i, \hat{\beta}), w(X_i, \hat{\beta})) = \\ \frac{1}{n(n-1)} \sum_{i=1}^n \left(g(X_i, V_i, D_i) w(X_i, \hat{\beta}) - \frac{\sum_{i=1}^n g(X_i, V_i, D_i) w(X_i, \hat{\beta})}{n} \right) \left(w(X_i, \hat{\beta}) - \frac{\sum_{i=1}^n w(X_i, \hat{\beta})}{n} \right),$$

where $cov(g(X_i, V_i, D_i) w(X_i, \hat{\beta}), w(X_i, \hat{\beta}))$ is 0 when $i \neq j$. The variance of Z is estimated as

$$\frac{1}{n} \sum_{i=1}^n \widehat{var}(w(X_i, \hat{\beta})) = \frac{1}{n(n-1)} \sum_{i=1}^n \left(w(X_i, \hat{\beta}) - \frac{\sum_{i=1}^n w(X_i, \hat{\beta})}{n} \right)^2.$$

Moreover, we have the estimate of sample mean values for S and Z as

$$\hat{\mu}_S = \frac{\sum_{i=1}^n w(X_i, \hat{\beta}) g(X_i, V_i, D_i)}{n}, \quad \hat{\mu}_Z = \frac{\sum_{i=1}^n w(X_i, \hat{\beta})}{n}.$$

Denoting g_i for $g(X_i, V_i, D_i)$ for notation simplicity, the variance of the ratio can be computed as

$$\begin{aligned}
\widehat{var}(S/Z) =& \frac{n/(n-1)}{\left(\sum_{i=1}^n w(X_i,\hat\beta)\right)^2} \left\{ \sum_{i=1}^n \left(w(X_i,\hat\beta) g_i - \frac{\sum_{i=1}^n w(X_i,\hat\beta) g_i}{n} \right)^2 \right. \\
&+ \left(\frac{\sum_{i=1}^n w(X_i,\hat\beta) g_i}{\sum_{i=1}^n w(X_i,\hat\beta)} \right)^2 \left(w(X_i,\hat\beta) - \frac{\sum_{i=1}^n w(X_i,\hat\beta)}{n} \right)^2 \\
&\left. - 2 \frac{\sum_{i=1}^n w(X_i,\hat\beta) g_i}{\sum_{i=1}^n w(X_i,\hat\beta)} \sum_{i=1}^n \left(w(X_i,\hat\beta) g_i - \frac{\sum_{i=1}^n w(X_i,\hat\beta) g_i}{n} \right) \left(w(X_i,\hat\beta) - \frac{\sum_{i=1}^n w(X_i,\hat\beta)}{n} \right) \right\} \\
=& \frac{n}{(n-1)\left(\sum_{i=1}^n w(X_i,\hat\beta)\right)^2} \sum_{i=1}^n w^2(X_i,\hat\beta) \left(g_i - \frac{\sum_{i=1}^n w(X_i,\hat\beta) g_i}{\sum_{i=1}^n w(X_i,\hat\beta)} \right)^2,
\end{aligned}$$

where we arrive at the same expression as (3.24). We can use the asymptotic normality result together with the above approximation for ratio variance to construct the 95% asymptotic confidence interval for the ratio estimator.

3.3 Simple analytic example

To illustrate the performance of the validation measures, we provide a simple toy example with a single binary risk factor, $X = 0$ or $X = 1$. The risk model for disease developed on the training set has only two values, with r_0 the risk of disease for $X = 0$, and r_1 for $X = 1$. The validation set of verified participants is of size N and has fraction f with $X = 0$ and $1 - f$ with $X = 1$. Let d_0 denote the percent of the Nf members of the verified validation set with $X = 0$ who have the disease. In other words $d_0 = \frac{1}{Nf} \sum_{i=1}^{Nf} D_i$, where $D_i = 1$ if the ith individual in the sum has the disease and 0 otherwise and the sum ranges over the individuals in the verified validation set with $X = 0$. The proportion of the verified validation set with $X = 1$ who have the disease is similarly defined: $d_1 = \frac{1}{N(1-f)} \sum_{i=1}^{N(1-f)} D_i$.

Then the usual $\widehat{CIL}$ calculation in (3.2) becomes:

$$\begin{aligned}
\widehat{CIL} &= \frac{fN(r_0 - d_0) + (1-f)N(r_1 - d_1)}{N}) \\
&= f(r_0 - d_0) + (1-f)(r_1 - d_1),
\end{aligned} \tag{3.26}$$

which is a weighted average of calibration for those with $X = 0$ and $X = 1$, with weight as the fraction of the validation set with the risk factor. Expression (3.26) explains the phenomena of differing validation performance for the same risk tool, and supports arguments in the clinical model assessment literature that validation is both a property of the risk tool and the validation set [Vickers et al., 2010]. If a validation set is “lucky” in the sense of having only a small fraction of individuals with the risk factor for which the training risk model shows poor calibration, then validation does not appear as poor as for a different validation set with a higher proportion of such cases.

For discrimination, we exam the TPR and FPR. Towards this, we assume the risk of disease for $X = 1$ is greater than that for $X = 0$, i.e. $r_1 > r_0$. Taking $(r_1 + r_0)/2$ as the threshold c, the usual $\widehat{TPR}(c)$ in (3.5) is the proportion of participants with disease and $X = 1$, i.e. the corresponding disease risk $r_1 > c$, among participants with disease. The $\widehat{FPR}(c)$ is then the proportion of participants without disease but with $X = 1$ among participants without disease. Hence, the usual $\widehat{TPR}(c)$ and $\widehat{FPR}(c)$ are

$$\widehat{TPR}(c) = \frac{(1-f)d_1}{(1-f)d_1 + fd_0}, \tag{3.27}$$

$$\widehat{FPR}(c) = \frac{(1-f)(1-d_1)}{(1-f)(1-d_1) + f(1-d_0)}. \tag{3.28}$$

To explore the performance of CIL_{MAR} adjusting for verification bias, we now let N include all participants regardless of being verified or not in the validation cohort. Let $\tilde{f}$ be the proportion among N with $X = 0$ ($1 - \tilde{f}$ for $X = 1$) in the validation cohort. We inherit the notations used in the above illustration example with all individuals being verified: r_0 (r_1) be the cancer risk given by the training cohort for $X = 0$ ($X = 1$), d_0 (d_1) be the probability of having cancer when $X = 0$ ($X = 1$) in the validation cohort, and p_0 (p_1) be the probability of being verified in the training when $X = 0$ ($X = 1$). We further denote $\tilde{p}_0$ ($\tilde{p}_1$) as the probability of being verified in the validation for $X = 0$ ($X = 1$). The estimated $\widehat{CIL}_{MAR}$ is then

$$\begin{aligned}\widehat{CIL}_{MAR} &= (1-\tilde{f})\tilde{p}_1 r_1 + \tilde{f}\tilde{p}_0 r_0 - \Big[(1-\tilde{f})\tilde{p}_1 d_1 + \tilde{f}\tilde{p}_0 d_0\Big] \\ &+ (1-\tilde{f})(1-\tilde{p}_1)r_1 - (1-\tilde{f})(1-\tilde{p}_1)d_1 + \tilde{f}(1-\tilde{p}_0)r_0 - \tilde{f}(1-\tilde{p}_0)d_0 \\ &= \tilde{f}(r_0 - d_0) + (1-\tilde{f})(r_1 - d_1). \end{aligned} \tag{3.29}$$

The $\widehat{CIL}_{MAR}$ has similar structure as $\widehat{CIL}$ just using different values to weight the calibration from individuals with $X = 0$ and $X = 1$. Therefore, the estimated $\widehat{CIL}$ and $\widehat{CIL}_{MAR}$ depends on the configuration of the training and validation population, such that if the weights are smaller for the subgroup calibrates poorly while higher for the well calibrated subgroup, the resulting CILs would close to 0.

Similarly, we have the verification bias adjusted $\widehat{TPR}_{MAR}(c)$ and $\widehat{FPR}_{MAR}(c)$ under MAR assumption using $c = (r_1 + r_0)/2$ as the threshold:

$$\widehat{TPR}_{MAR}(c) = \frac{(1-\tilde{f})d_1}{(1-\tilde{f})d_1 + \tilde{f}d_0}, \tag{3.30}$$

$$\widehat{FPR}_{MAR}(c) = \frac{(1-\tilde{f})(1-d_1)}{(1-\tilde{f})(1-d_1) + \tilde{f}(1-d_0)}. \tag{3.31}$$

Comparing the TPR_{MAR} and FPR_{MAR} with their corresponding usual metrics, we have

$$\widehat{TPR}(c) - \widehat{TPR}_{MAR}(c) = \frac{(\tilde{f} - f)d_1 d_0}{[(1-f)d_1 + fd_0][(1-\tilde{f})d_1 + \tilde{f}d_0]},$$
$$\widehat{FPR}(c) - \widehat{FPR}_{MAR}(c) = \frac{(\tilde{f} - f)(1-d_1)(1-d_0)}{[(1-f)(1-d_1) + f(1-d_0)][(1-\tilde{f})(1-d_1) + \tilde{f}(1-d_0)]}.$$

The $\widehat{TPR}_{MAR}$ and $\widehat{FPR}_{MAR}$ will change in the same direction compared to $\widehat{TPR}$ and $\widehat{FPR}$, respectively. Either both increase or both decrease, guiding by $\tilde{f} - f$. Hence, the improvement in the corresponding AUC_{MAR} compared to AUC are not guaranteed.

Numerical exploration

Following the illustration of analytic example before, we explore the it numerically using only one risk factor X follows the binomial distribution in this part. For the training cohort, we let $X \sim Binomial(0.7)$, the verification probabilities given X be $(P(V = 1|X = 0), P(V = 1|X = 1)) = (0.3, 0.6)$, and the cancer probabilities given X be $(P(D = 1|X = 0), P(D = 1|X = 1)) = (0.2, 0.8)$. Using notation in the analytic example section, we have $r_0 = 0.2$ and $r_1 = 0.8$ in this case. When X, V, and D in the validation cohort are generated exactly as the training cohort, we obtain $d_0 = r_0$ and $d_1 = r_1$ so that $\widehat{CIL}$ is close to 0. If in the validation cohort, $(P(D = 1|X = 0), P(D = 1|X = 1)) = (0.2, 0.5)$. The percentage of the verified participants who have cancer and $X = 1$ among those verified participants with $X = 1$ in the validation cohort, i.e. d_1, is 0.5 versus $d_0 = 0.2$. The estimated $\widehat{CIL}$ is

$$\widehat{CIL} = \frac{0.3 \times 0.3}{0.3 \times 0.3 + 0.7 \times 0.6}(0.2 - 0.2) + \left(1 - \frac{0.3 \times 0.3}{0.3 \times 0.3 + 0.7 \times 0.6}\right)(0.8 - 0.5) \approx 0.247. \tag{3.32}$$

For the TPRs and FPRs using only verified participants from the validation cohorts, we have

$$\widehat{TPR} = \frac{(1 - 0.9/0.51) \times 0.5}{(1 - 0.9/0.51) \times 0.5 + 0.9/0.51 \times 0.2} \approx 0.921$$
$$\widehat{FPR} = \frac{(1 - 0.9/0.51) \times (1 - 0.5)}{(1 - 0.9/0.51) \times (1 - 0.5) + 0.9/0.51 \times (1 - 0.2)} \approx 0.745.$$

To incorporate the adjustment for verification bias, we let the proportion with $X = 0$ in the validation set be 0.3, i.e. $\tilde{f} = 0.3$. We use $(d_1, d_0) = (0.5, 0.2)$ and assume that the probability of being verified in the validation cohort for $X = 0$ $(X = 1)$, $\tilde{p}_0 = p_0 = 0.3$ $(\tilde{p}_1 = p_1 = 0.6)$, i.e. the same as those from the training cohort. The estimated $\widehat{CIL}_{MAR}$ is

$$\widehat{CIL}_{MAR} = (1 - \tilde{f})(r_1 - d_1) = 0.7 \times (0.8 - 0.5) = 0.21. \tag{3.33}$$

We weight the subgroup with $X = 1$ by 0.7, while around 0.8 for $\widehat{CIL}$. Therefore, the $\widehat{CIL}_{MAR}$ is slightly improved compared to $\widehat{CIL}$.

Similarly, we have

$$\widehat{TPR}_{MAR} = \frac{(1-0.3) \times 0.5}{(1-0.3) \times 0.5 + 0.3 \times 0.2} \approx 0.854,$$
$$\widehat{FPR}_{MAR} = \frac{(1-0.3) \times (1-0.5)}{(1-0.3) \times (1-0.5) + 0.3 \times (1-0.2)} =\approx 0.593.$$

3.4 Application

To illustrate the methods with real data, we develop and validate a risk tool to predict prostate cancer with data from PLCO and SELECT. For training a prostate cancer risk prediction model, individual-level level data from PLCO are used [Andriole et al., 2009]. For validating the PLCO risk model, individual-level data from the SELECT are used [Lippman et al., 2009]. Table 2.1 provides a contrast the two cohorts in risk factors. Compared to the PLCO training set, the SELECT validation set is elder with more African-Americans, has a lower rate of abnormal digital rectal exams (DRE), and higher rates of first-degree family histories or prostate cancer and prior negative biopsies. The rates of verification by biopsy in both cohorts are similar at approximately 12.7%.

We perform exploratory graphical analysis of the differential verification in the PLCO training and SELECT validation sets in Figure 2.2. Protocols of both trials recommended biopsy for prostate-specific antigen (PSA) $>$ 4 ng/mL or abnormal DRE or both, hence the high odds ratios for verification for these risk factors. However, a PLCO participant with PSA $>$ 4 ng/mL has twice the odds of receiving a biopsy compared to a SELECT participant, whereas the reverse is true for a participant with abnormal DRE. Odds ratios for the other risk factors do not differ as much between the cohorts.

The vector of risk factors collected for all participants in both trials are $X = (PSA, DRE,$ $age, African\ ancestry, family\ history, prior\ negative\ biopsy)$, with the last three variables as binary indicators, and family history indicating prostate cancer first-degree relative history. Transformations of PSA are used when helpful for improving the fit of models including the log-base-2 transformation ($log_2 PSA$) and an indicator that PSA exceeds 4 ng/mL ($I(PSA > 4)$). In the model for cancer risk, we use either of these two transformed PSA values. As both studies have a longitudinal follow-up, we only use the last measurements of the risk factors in the analysis, with designation as last depending on verification and cancer outcome. For men who never receive a biopsy during the trial ($V = 0, D = missing$), the last recorded risk factors in the study are used. For men with multiple negative biopsies ($V = 1, D = 0$), the last biopsy in the study with a PSA value within two years before it is used. For men who ever have a positive biopsy, only the first positive biopsy is used subject to having a PSA value within two years prior ($V = 1, D = 1$). When multiple PSA measurements are available within two years before biopsy, the one most recent to biopsy is used.

We fit a prostate cancer risk prediction model $R(X)$ by applying a multivariable logistic regression to model the outcome of prostate cancer on the 3813 PLCO participants who received a biopsy, which model form is commonly used in modeling the risk of a dichotomous event. Odds ratios for prostate cancer from the model are shown in Table 3.1. The PLCO prostate cancer risk model indicates that higher PSA, African ancestry, having a family history of prostate cancer, and no prior negative biopsy are predictive of a higher risk of prostate cancer. Particularly, the main effect of abnormal DRE is not included in the model described in Table 3.1 since including it gives an estimated odds ratio smaller than 1 which may due to its interaction term with log-base-2 PSA (See Table 3.3). From the model in Table 3.1, when the PSA value increases, the cancer risk for men with abnormal DRE will increase faster than those with normal DRE as shown in Figure 3.2, which applies the estimated model to participants with various PSA and DRE values. To account for the verification bias within the validation cohort, we fit a logistic model for the validation cancer risk in the validation cohort, which model is shown in Table 3.2 and whose predictions are used to impute the missing cancer outcomes for the unverified participants in the validation cohort.

Table 3.1: Odds ratios for prostate cancer from logistic regression built on 3813 PLCO participants verified by biopsy,1833 (48.1%) of which had prostate cancer; All p-values are significant and less than 0.001 except for $African\ ancestry$ and $Family\ history$ (both p-values = 0.01). log_2PSA = log-base-2 of prostate-specific antigen; DRE = digital rectal examination.

Risk factors	Odds ratio	95% confidence interval
$Intercept$	0.20	(0.17, 0.24)
log_2PSA	1.92	(1.80, 2.06)
$African\ ancestry$	1.50	(1.10, 2.05)
$Family\ history$	1.39	(1.10, 1.74)
$Prior\ negative\ biopsy$	0.52	(0.42, 0.65)
$log_2PSA * DRE$	1.24	(1.16, 1.33)

Calibration-in-the-large estimate (95% confidence intervals) for the PLCO risk model evaluated on the verified SELECT participants is calculated as $\widehat{CIL} = -0.056\ (-0.070, -0.042)$,

Table 3.2: Odds ratios for prostate cancer from logistic regression built on 4185 SELECT participants verified by biopsy, 2028 (48.5%) of which had prostate cancer; All p-values are significant and less than 0.001. log_2PSA = log-base-2 of prostate-specific antigen; DRE = digital rectal examination.

Risk factors	Odds ratio	95% confidence interval
$Intercept$	0.09	(0.04, 0.18)
log_2PSA	1.91	(1.77, 2.06)
$DRE(abnormal)$	2.30	(1.96, 2,72)
Age	1.02	(1.00, 1.03)
$Family\ history$	1.64	(1.42, 1.91)
$Prior\ negative\ biopsy$	0.54	(0.45, 0.66)

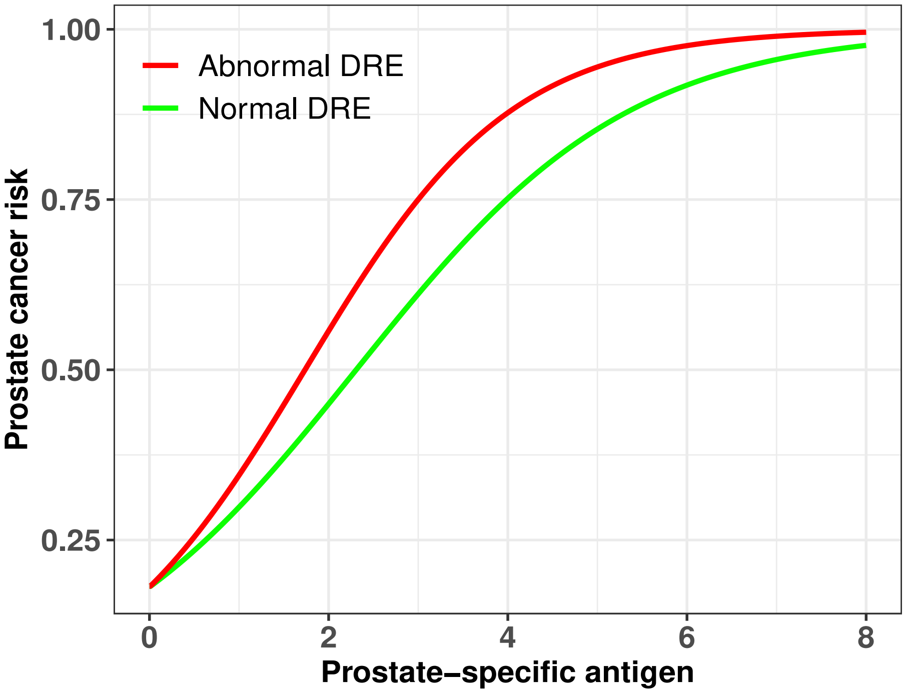

Figure 3.2: Estimated prostate cancer risk from the risk model built on 3813 verified PLCO participants in Table 3.1 with various prostate-specific antigen and digital rectal exam (DRE) values with family history of prostate cancer ($Family\ history = 1$), prior negative biopsy ($Prior\ negative\ biopsy = 1$), and African ancestry ($African\ ancestry = 1$).

Table 3.3: Odds ratios for prostate cancer from logistic regression including main effect of binary DRE built on 3813 PLCO participants verified by biopsy,1833 (48.1%) of which had prostate cancer. All p-values are significant and less than 0.001 except for $intercept$, $African\ ancestry$, and $Family\ history$ (p-values: 0.005, 0.021, and 0.006, respectively). log_2PSA = log-base-2 of prostate-specific antigen; DRE = digital rectal examination.

Risk factors	Odds ratio	95% confidence interval
$Intercept$	0.64	(0.46, 0.87)
log_2PSA	1.29	(1.15, 1.45)
$DRE(abnormal)$	0.22	(0.15, 0.33)
$African\ ancestry$	1.45	(1.06, 1.99)
$Family\ history$	1.38	(1.10 1.74)
$Prior\ negative\ biopsy$	0.54	(0.44, 0.68)
$log_2PSA * DRE$	2.16	(1.84, 2.53)

indicating under-prediction of the PLCO model for SELECT since the value is below zero. When taking all SELECT participants into account under the missing-at-random assumption, the estimated $\widehat{CIL}_{MAR} = -0.029\ (-0.051, -0.009)$ is better than the $\widehat{CIL}$ calculated with only the verified participants. The estimated $\widehat{CIL}_{MAR}$ has a wider confidence interval compared to the $\widehat{CIL}$ based on verified participants, which does not surprise us since metrics involve more participants should have higher variation. The AUC showing the discrimination of the PLCO risk model on SELECT is $\widehat{AUC} = 0.674\ (0.661, 0.690)$ based on verified participants in SELECT, while $\widehat{AUC}_{MAR} = 0.718\ (0.699, 0.741)$ using all participants. We can see from Figure 3.3 that the gap between TPR and FPR is smaller than that between TPR_{MAR} and FPR_{MAR} explaining the higher value of $\widehat{AUC}_{MAR}$ compared to the usual estimate.

Table 3.4: Estimated calibration-in-the-large (CIL) and area-under-the-receiver-operating-characteristic curve (AUC) among 4185 verified or 32629 SELECT participants. Estimates with subscript MAR are calculated from 32629 SELECT participants under missing-at-random (MAR) assumption, while from 4185 verified participants for the others. 95% confidence intervals are from bootstrapping with 600 repetitions.

	Estimate	95% confidence interval
CIL	-0.056	(-0.070, -0.041)
CIL_{MAR}	-0.029	(-0.052, -0.007)
AUC	0.674	(0.659, 0.690)
AUC_{MAR}	0.718	(0.699, 0.742)

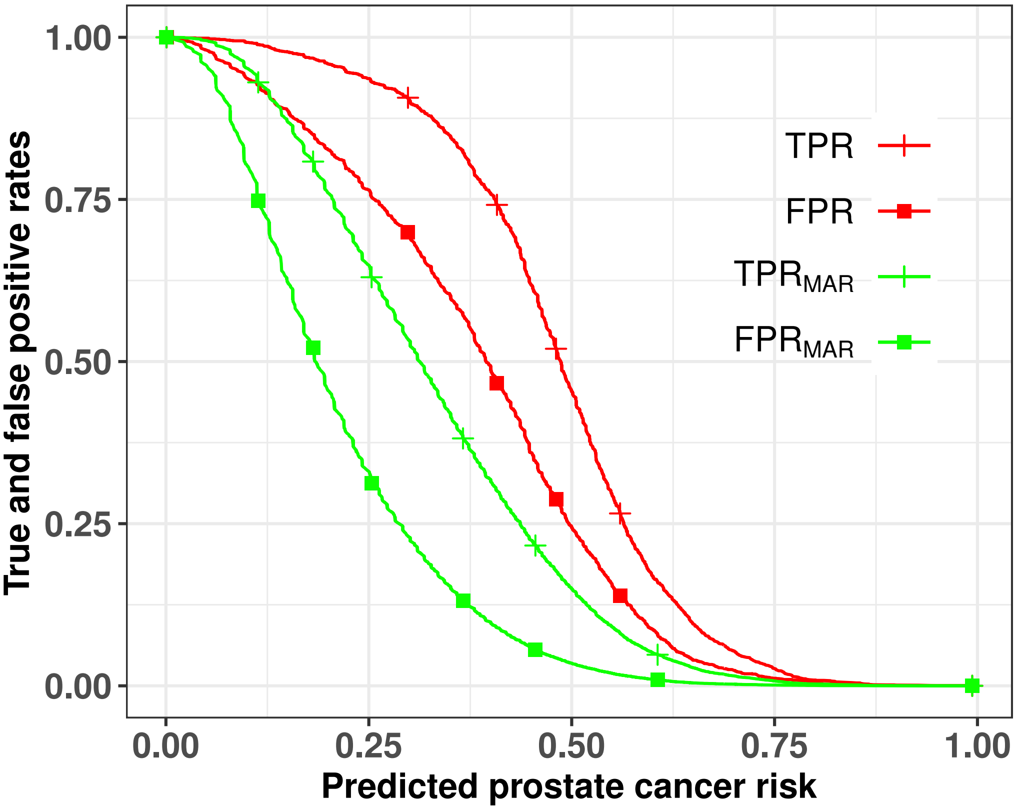

Figure 3.3: True positive rates (TPRs) and false positive rates (FPRs) calculated among 4185 verified or 32629 SELECT participants, where the latter is used for measures with MAR in the subscript under the missing-at-random (MAR) assumption.

3.5 Summary

In this chapter, we reviewed the common measures used in the external validation of clinical risk prediction models, where the calibration-in-the-large (CIL) for evaluating the calibration and the area under the receiver operating characteristic (AUC) for assessing the discrimination were examined. The calculation of the usual CIL and AUC on the validation cohort is based on the verified participants only, while ignoring those who have not been verified and therefore, without known disease outcomes. However, the verified participants may not be a random sample of the cohort but rather those who meet certain criteria, such as only those who have PSA level $>$ 4 ng/mL and suspicious DRE will be referred to further biopsy to ascertain prostate cancer. Hence, the characteristics of the verified participants may differ from that of the unverified ones even if they belong to the same cohort. Such difference is often referred to as the verification bias [Begg and Greenes, 1983; Alonzo and Pepe, 2005]. To adjust for such verification bias arising from ignoring the unverified participants in the validation, we imputed their missing outcome with the risk of having the disease as proposed by Alonzo and Pepe [2005] under the missing-at-random assumption, where the risk of disease is given by the risk model built on the verified participants in the validation cohort. Moreover, we also showed the consistency and asymptotic normality properties of the estimators in this chapter.

We illustrated the method with data from two large North American prostate cancer screening and prevention trials, i.e. the Selenium and Vitamin E Cancer Prevention Trial (SELECT) and the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. In the application, the estimated $\widehat{CIL}_{MAR}$ was closer to the ideal 0 compared to $\widehat{CIL}$, while $\widehat{AUC}_{MAR}$ was higher than $\widehat{AUC}$. The estimates accounting for verification bias had wider confidence intervals and hence, larger variations, compared to the usual metrics since they involved more participants.

4 A novel external validation method for adjusting for training and validation cohort differences

In this chapter, we propose a novel weighted version of the validation metrics that accommodates the differences in the risk factor distributions and in the outcome verification procedures between the training and validation sets, which provides a more comprehensive assessment of model performance when individual level information from both the training and validation sets is available. We specifically formalize the concepts of reproducibility and transportability when externally validating clinical models taking the accommodation of the impact of heterogeneity of risk factor distributions and verification procedures into account. Towards this, the calibration and discrimination of a model are evaluated. We illustrate the method with 29699 PLCO and 26422 SELECT participants shown in Section 2.3.

4.1 Research context

Once a clinical risk model has been built on the training cohort, we then validate it internally or externally. For the former procedure, i.e. the internal validation, the cross-validation method is often used [Steyerberg, 2019]. In the cross-validation, a subset will be sampled randomly from the training data for the prediction model and used to validate the built model, which can be viewed as assessing the reproducibility of the model because the validated sample is from the same population as the training data. On the other hand, external validation can provide a more rigorous assessment of model performance than internal validation only, which procedure uses external samples from different populations other than the training one and is often viewed as assessing the transportability of the model because it involves samples with different characteristics compared to the training set [Debray et al., 2015].

In the external validation, researchers have recognized that the heterogeneity between the training and external validation cohorts may affect the performance of the prediction model, which heterogeneity is also known as “case-mix differences” [Austin et al., 2016; Steyerberg, 2019; Davis et al., 2019; Song et al., 2020]. Most efforts to address this issue are qualitative or descriptive, such as comparisons of patient characteristics between the training and validation data sets [Debray et al., 2015; Ankerst et al., 2012; Metnitz et al., 2000; Tolksdorf et al., 2019]. Other approaches include refitting the training model on the validation data and simulating outcomes in the validation data under the assumption that the training model is correct [Vergouwe et al., 2010]. This approach can be used to determine whether the originally observed performance is much worse than what would have been observed in the validation data in the ideal case. Powers et al. [2019] weighted observations from the training data to see how the model would perform in a target population with a different distribution of risk factors, but they did not use weighting to adjust such difference in the calculation of model performance metrics.

However, a model that performs poorly and fails to transport to an external population with different distributions of risk factors or different methods of referral for definitive disease diagnosis may still be valuable. Here, we refer to model failure because the external population has different risk factor distributions as “selection bias”, while model failure because the external population uses different procedures to refer patients for definitive diagnoses as “verification bias”. Such differences in risk factor distributions and verification procedures typically reflect true differences in populations.

In a clinical trial, we only know the status of the event of interest for participants who have been verified, but not for the unverified ones. The distribution of the event of interest depends on the verification procedures, i.e. evaluation of diagnostic tests. Ideally, all participants in the trial regardless of having diseased or not should be verified or random samples from the trial population would be verified and hence, no verification bias occurs in this case. However, in practice, only certain participants may be referred to receive the verification test to assure the disease status based on their risk factors. For example, it might be desirable to verify all men with or without prostate cancer to determine the effect of a new substance in a clinical trial. But physicians may decide who will be biopsied based on the characteristics of the man such as elevated PSA level and abnormal DRE and hence, verification bias presents. Begg and Greenes [1983] and Alonzo and Pepe [2005] developed weighting methods to correct for non-random selection for verification, i.e. correction of the verification bias. If a risk model has been developed in a training population in which verification mechanisms differ from those in the external validation population, verification bias occurs and can impair risk model performance in the external population. Here, we use weighting to account for differences in disease verification with a different objective than that in Begg and Greenes [1983] and Alonzo and Pepe [2005], where we focus on eliminating the verification bias from different verification mechanisms used in different populations. When only certain participants in a clinical trial have been verified but not all, the disease risk prediction model built on the data of the verified participants does not predict disease risk, but rather predicts the joint probability that the participant is verified, i.e. receives the diagnostic test, and such diagnostic test shows positive for the disease. Thus, we design a weighting procedure to see how well the risk model would predict this joint probability if the external population has similar verification procedures as the training population.

Here, we focus on external validation of the risk prediction model and propose a novel weighting procedure to adjust for the selection and verification bias in the calculation of the measures assessing model performance, such as calibration and discrimination. By comparing unadjusted and adjusted measures of performance, we can gain insight into which factors might contribute to mis-calibration in the external population. In particular, we also give the conditions on the risk model and the characteristics of training and validation populations that ensure reproducibility or transportability of the model on the external population, where we check these conditions with unweighted versus the proposed weighted performance measures.

In the following Section 4.2, we introduce notation, define the performance metrics, develop the adjustment method, and present the estimates of the weighted performance metrics and their asymptotic properties. Section 4.3 evaluates the robustness of these methods when assumptions violate via simulations. In Section 4.4, we illustrate our method by building and validating a survival model that predicts prostate cancer incidence using two large North American prostate cancer screening trials. The two trial populations differ not only in their distributions of risk factors but also in their PSA screening scheme that use to decide the timing of biopsies for prostate cancer diagnosis (see Section 2.3 for more information). We close with a brief discussion of our approach in Section 4.5.

4.2 Method

Notation and assumptions

Let $\mathbf{X}$ denote the vector of model predictors. We assume that a risk model $R = R(\mathbf{X})$ estimates the probability of an outcome $D = 1$ for those who have the outcome of interest and $D = 0$ for those who do not. Let N_T, $T = 0, 1$, be the sizes of the validation (N_0) and training (N_1) data sets. We assume that the risk model R is estimated from a sample of a training population $(T = 1)$. The performance of R will be assessed in a sample from an independent validation population ($T = 0$). R can be a logistic model if we predict disease prevalence, or it can be an absolute risk from a survival model that predicts disease incidence over a given projection period τ. We assume there are additional risk factors/-covariates $\mathbf{Z}$ that are not included in the prediction model R but are available in both the training and validation sets, and let $\mathbf{X}^* = (\mathbf{X}, \mathbf{Z})$.

For each individual in the validation set, we compute the risk estimate $R_i = R(\mathbf{X}_i)$ given the respective risk factor $\mathbf{X}_i$, for $i = 1, \ldots, N_0$. We will add the verification mechanism later. For models that depend on a projection period τ, we assume that we observe D up to the end of the τ follow-up period. We will address the censoring of outcomes in a later section. For notation brevity, we omit τ in the following and denote the validation data by $(R_i, D_i, \mathbf{X}_i^*), i = 1, \ldots, N_0$.

We assume that the true probabilities of outcome in the training and validation populations are $\pi_T(\mathbf{X}^*) = P(D = 1|\mathbf{X}^*, T),\ T = 0, 1$. We assume the disease status D depends only on predictors $\mathbf{X}$, i.e.

$$\pi_T(\mathbf{X}) = P(D = 1|\mathbf{X}, T) = \int_{\mathbf{z}} \pi_T(\mathbf{X}, \mathbf{z}) dF_T(\mathbf{z}|\mathbf{X}), \quad T = 0, 1. \tag{4.1}$$

Here, $\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$ implies $\pi_0(\mathbf{X}) = \pi_1(\mathbf{X})$ only if $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$ holds in addition, i.e. the conditional distributions of the omitted risk factor ($\mathbf{Z}$) given the predictors $\mathbf{X}$ used in modeling R are the same between the two data sets.

Performance metrics

We discuss the model calibration first followed by the measures of accuracy and discrimination, where the former refers to evaluating if the model predictions and observations are identical, while the latter is about assessing if the model can separate disease versus non-disease participants correctly.

Calibration measures

We define a model R to be *strongly calibrated* if $P(D = 1|\mathbf{X} = \mathbf{x}) = R(\mathbf{x})$ for all values of $\mathbf{X}$, while *well calibrated* if the predicted and observed numbers of events agree in the subsets or the overall population, i.e.

$$C = \frac{E(D)}{E(R)} = \frac{\int_{\mathbf{x}^*} \pi(\mathbf{x}^*) dF(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x}) dF(\mathbf{x})} = 1, \tag{4.2}$$

where $F(\mathbf{x})$ and $F(\mathbf{x}^*)$ denote the distributions of $\mathbf{X}$ and $\mathbf{X}^*$, respectively. Obviously, if the model is strongly calibrated, $P(D = 1) = E[R(\mathbf{X})]$ and therefore, R is also well calibrated [Pfeiffer and Gail, 2017]. We focus on the overall calibration for the ease of exposition in this section and discuss calibration in subgroups in more detail in the data example.

From the definition in (4.2), a model R that is strongly calibrated in the training data ($T = 1$) can only be strongly calibrated in the validation data ($T = 0$) as well if $\pi_0(\mathbf{x}) = \pi_1(\mathbf{x}) = R(\mathbf{x})$. This condition holds when $\pi_1(\mathbf{x}^*) = \pi_0(\mathbf{x}^*)$ and $F_0(\mathbf{z}|\mathbf{x}) = F_1(\mathbf{z}|\mathbf{x})$, i.e. given these two equations and (4.1):

$$P_0(D = 1|\mathbf{X}) \quad = \quad \int_{\mathbf{z}} \pi_0(\mathbf{X}, \mathbf{z}) dF_0(\mathbf{z}|\mathbf{X}) = \int_{\mathbf{z}} \pi_1(\mathbf{X}, \mathbf{z}) dF_1(\mathbf{z}|\mathbf{X}) = \pi_1(\mathbf{X}) = R(\mathbf{X}).$$

The difference between predication and observations, i.e. the calibration-in-the-large (CIL), is an alternative measure to assess calibration:

$$CIL = E(R(\mathbf{X}) - D). \tag{4.3}$$

When $C = 1$, we have$E(D) = E[R(\mathbf{X})]$ and thus $CIL = E(R(\mathbf{X}) - D)$. Therefore, a test for $C = 1$ is equal to testing $CIL = 0$ and we focus on C here.

The calibration ratio on the validation cohort can be estimated consistently by replacing the expectation in (4.2) by the empirical mean as

$$\widehat{C}_0 = \frac{\sum_{i=1}^{N_0} D_i}{\sum_{i=1}^{N_0} R(\mathbf{X}_i)}. \tag{4.4}$$

Accuracy and discrimination measures

Commonly used accuracy measures for clinical decision making, such as recommending a patient for further clinical practice, depending on a particular risk threshold r^* are the true positive rate (TPR) and false positive rate (FPR), i.e.

$$TPR_0(r^*) = P_0(R \geq r^*|D = 1) = \frac{E_0\{I(R(\mathbf{X}) \geq r^*)D\}}{E_0(D)}, \tag{4.5}$$

and

$$FPR_0(r^*) = P_0(R \geq r^*|D = 0) = \frac{E_0\{I(R(\mathbf{X}) \geq r^*)(1-D)\}}{E_0(1-D)}, \tag{4.6}$$

where I denotes the indicator function and subscript 0 denotes for validation set. TPR and FPR are estimated consistently by

$$\widehat{TPR}_0(r^*) = \frac{\sum_{i=1}^{N_0} I(R(\mathbf{X}_i) \geq r^*)D_i}{\sum_{i=1}^{N_0} D_i} \tag{4.7}$$

$$\widehat{FPR}_0(r^*) = \frac{\sum_{i=1}^{N_0} I(R(\mathbf{X}_i) \geq r^*)(1-D_i)}{\sum_{i=1}^{N_0}(1-D_i)}. \tag{4.8}$$

The receiver operator characteristic (ROC) curve plots the TPR on the y-axis against the FPR on the x-axis across different risk thresholds. A summary measure of the ROC curve evaluating the discrimination of the model is the area-under-the-receiver-operating-characteristic curve (AUC), which can be computed as the probability that a randomly selected case ($D_i = 1$) has a larger risk estimate than a randomly selected control ($D_j = 0$), $AUC_0 = P_0(R(\mathbf{X}_i) > R(\mathbf{X}_j)|D_i = 1, D_j = 0)$. The AUC on the validation set can be written in terms of the risk factor distribution in cases and non-cases by applying Bayes theorem to $F_0(\mathbf{x}|D), D = 0, 1$ as

$$AUC_0 = \frac{E_0\big[I\{R(\mathbf{X}) > R(\tilde{\mathbf{X}})\}\pi_0(\mathbf{X})\{1 - \pi_0(\tilde{\mathbf{X}})\}\big]}{E_0(D)\{1 - E_0(D)\}}. \tag{4.9}$$

An empirical estimate of (4.9) in the validation cohort with the accommodation of ties is

$$\widehat{AUC}_0 = \frac{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i(1-D_j)[I(R(\mathbf{X}_i) > R(\mathbf{X}_j) + 0.5I(R(\mathbf{X}_i) = R(\mathbf{X}_j))]}{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i(1-D_j)}. \tag{4.10}$$

Accommodating selection bias

We first define the weights addressing the differences in the distributions of risk factor $\mathbf{X}^*$ between training and validation sets, i.e. accounting for the "selection bias", and present the impact of such differences with the weighted and unweighted validation performance measures afterward.

Selection weighted performance measures

Towards accommodating the difference in the distribution of predictors between training and validation sets, we propose the *selection weights*:

$$w(\mathbf{X}^*) = \frac{dF_1(\mathbf{X}^*)}{dF_0(\mathbf{X}^*)} = \frac{P(\mathbf{X}^*|T=1)}{P(\mathbf{X}^*|T=0)} = \frac{P(T=1|\mathbf{X}^*)P(T=0)}{P(T=0|\mathbf{X}^*)P(T=1)}. \tag{4.11}$$

The corresponding *selection weighted calibration ratio* on validation cohort is

$$C_0^W = \frac{E_0[Dw(\mathbf{X}^*)]}{E_0[R(\mathbf{X})w(\mathbf{X}^*)]} = \frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)w(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*} R(\mathbf{x})w(\mathbf{x}^*)dF_0(\mathbf{x}^*)}. \tag{4.12}$$

Here, we weight both the predictions and the observations adjusting for the difference in the distribution of risk factors between the training and validation populations. The proposed weights can depend on a broader set of variables $\mathbf{X}^*$, i.e. including additional variables that might impact the true probability of the outcome (π) other than the model predictors.

The *selection weighted* TPR and FPR on validation cohort are

$$TPR_0^W(r^*) = \frac{E_0\{w(\mathbf{X}^*)I(R(\mathbf{X}) \geq r^*)D\}}{E_0\{w(\mathbf{X}^*)D\}}, \tag{4.13}$$

$$FPR_0^W(r^*) = \frac{E_0\{w(\mathbf{X}^*)I(R(\mathbf{X}) \geq r^*)(1-D)\}}{E_0\{w(\mathbf{X}^*)(1-D)\}}. \tag{4.14}$$

The *selection weighted* AUC is defined as

$$AUC_0^W = \frac{E_0\big[I\{R(\mathbf{X}) > R(\tilde{\mathbf{X}})\}\pi_0(\mathbf{X})w(\mathbf{X}^*)\{1-\pi_0(\tilde{\mathbf{X}})\}w(\tilde{\mathbf{X}}^*)\big]}{E_0\{Dw(\mathbf{X}^*)\}E_0\{(1-D)w(\mathbf{X}^*))\}}. \tag{4.15}$$

Properties of unweighted and selection weighted performance measures in the validation cohort

We assume the model is well calibrated in the training cohort, i.e.

$$C_1 = \frac{E_1(D)}{E_1\{R(\mathbf{X})\}} = \frac{\int_{\mathbf{x}^*} \pi_1(\mathbf{x}^*) dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x}) dF_1(\mathbf{x})} = 1, \tag{4.16}$$

and the true disease probabilities between the training and validation populations are the same, i.e.

$$\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*). \tag{4.17}$$

Model reproducibility

In addition to $\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$, we assume that the predictor distributions and the conditional covariate distributions are the same between the two populations, i.e. $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$ and $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$, respectively. These two additional assumptions imply that $F_0(\mathbf{X}) = F_1(\mathbf{X})$ because $F_T(\mathbf{X}) = F_T(\mathbf{X}^*)/F_T(\mathbf{Z}|\mathbf{X})$ for $T = 0, 1$ by Bayes theorem. In this situation, the validation and training populations do not differ in disease probabilities and distributions of risk factors that are relevant for assessing the performance of the prediction model R. The weights in (4.11) are all one i.e. $w(\mathbf{X}^*) = 1$, since $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$. Therefore, the unweighted and weighted calibration measures are the same,

$$C_0 = \frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*) dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x}) dF_0(\mathbf{x})} = \frac{\int_{\mathbf{x}^*} \pi_1(\mathbf{x}^*) dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x}) dF_1(\mathbf{x})} = C_0^W = C_1 = 1. \tag{4.18}$$

In this situation, the risk model is “reproducible” as it gives identical calibration results in the training and validation populations [Debray et al., 2015].

Actually, the model is reproducible as long as

$$\int_{\mathbf{x}^*} \pi_1(\mathbf{x}^*) w(\mathbf{x}^*) dF_1(\mathbf{x}^*) = \int_{\mathbf{x}} R(\mathbf{x}) dF_1(\mathbf{x}), \tag{4.19}$$

i.e. R is well calibrated in the training population. But it does not require $R(\mathbf{X}) = \pi_1(\mathbf{X})$ for each $\mathbf{x}$ value, which is rather referred to as R is strongly calibrated.

Model transportability

We now still assume $\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$, but weaken the assumptions made on the risk factor distribution by only assuming $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$, but allowing $F_0(\mathbf{X}) \neq F_1(\mathbf{X})$. Then, we have

$$\pi_0(\mathbf{X}) = \int_{\mathbf{z}} \pi_0(\mathbf{X}, \mathbf{z}) dF_0(\mathbf{z}|\mathbf{X}) = \int_{\mathbf{z}} \pi_1(\mathbf{X}, \mathbf{z}) dF_1(\mathbf{z}|\mathbf{X}) = \pi_1(\mathbf{X}),$$

according to the definition of true disease probability (4.1). If we assume R is strongly calibrated in the training data $(T = 1)$ in addition, i.e. $\pi_1(\mathbf{x}) = P_1(D = 1|\mathbf{X} = \mathbf{x}) = R(\mathbf{x})$ for any $\mathbf{x}$, we have

$$C_0 = \frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*) dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x}) dF_0(\mathbf{x})} = \frac{\int_{\mathbf{x}} \pi_0(\mathbf{x}) dF_0(\mathbf{x})}{\int_{\mathbf{x}} R(\mathbf{x}) dF_0(\mathbf{x})} = \frac{\int_{\mathbf{x}} \pi_1(\mathbf{x}) dF_0(\mathbf{x})}{\int_{\mathbf{x}} R(\mathbf{x}) dF_0(\mathbf{x})} = 1, \tag{4.20}$$

i.e. the model is "transportable" as the unweighted calibration ratio on validation cohort is equal to the ideal one [Debray et al., 2015]. In the second equality of (4.20), we replace the numerator $\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)dF_0(\mathbf{x}^*)$ with $\int_{\mathbf{x}} \pi_0(\mathbf{x})dF_0(\mathbf{x})$ because in discrete case, we have

$$\begin{aligned}
& \sum_{\mathbf{x}}\sum_{\mathbf{z}} \pi_0(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})P_0(\mathbf{z},\mathbf{x}) \\
= & \sum_{\mathbf{x}}\sum_{\mathbf{z}} \pi_0(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})P_0(\mathbf{z}|\mathbf{x})P_0(\mathbf{x}) \\
= & \sum_{\mathbf{x}}\sum_{\mathbf{z}} \frac{P_0(D=1,\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}{P_0(\mathbf{X}=\mathbf{x},\mathbf{Z}=\mathbf{z})}\frac{P_0(\mathbf{z},\mathbf{x})}{P_0(\mathbf{x})}P_0(\mathbf{x}) \\
= & \sum_{\mathbf{x}}\left[\sum_{\mathbf{z}} P_0(D=1,\mathbf{Z}=\mathbf{z}|\mathbf{X}=\mathbf{x})\right]P_0(\mathbf{x}) \\
= & \sum_{\mathbf{x}} P_0(D=1|\mathbf{X}=\mathbf{x})P_0(\mathbf{x}) = \sum_{\mathbf{x}}\pi_0(\mathbf{x})P_0(\mathbf{x}).
\end{aligned}$$

The continuous case follows similarly.

If R is just well but not strongly calibrated in the training data, then the unweighted calibration ratio C_0 is not equal to C_1, i.e.

$$C_0 = \frac{\int_{\mathbf{x}^*}\pi_0(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_0(\mathbf{x})} = \frac{\int_{\mathbf{x}}\pi_1(\mathbf{x})dF_0(\mathbf{x})}{\int_{\mathbf{x}} R(\mathbf{x})dF_0(\mathbf{x})} \neq C_1 = 1 \tag{4.21}$$

since $\pi_1(\mathbf{x}) \neq R(\mathbf{x})$ for any $\mathbf{x}$ under well calibrated assumption. However, the weighted calibration ratio has

$$C_0^W = \frac{\int_{\mathbf{x}^*}\pi_0(\mathbf{x}^*)w(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*} R(\mathbf{x})w(\mathbf{x}^*)dF_0(\mathbf{x}^*)} = \frac{\int_{\mathbf{x}^*}\pi_0(\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_1(\mathbf{x})} = \frac{\int_{\mathbf{x}^*}\pi_1(\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_1(\mathbf{x})} = C_1 = 1, \tag{4.22}$$

provided the weights are correctly modeled.

Thus, the evaluation of the transportability of the risk predication model R corresponds to assessing that if $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$ and R is strongly calibrated by checking if $C_0 = 1$ or alternatively, if $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$ and R is well calibrated by checking if $C_0 \neq 1$ but $C_0^W = 1$.

Similarly in the assessment for accuracy and discrimination, we have $TPR_0(r^*) = TPR_1(r^*)$, $FPR_0(r^*) = FPR_1(r^*)$ and $AUC_0 = AUC_1$ when $\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$, $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$ and R is strongly calibrated. Whereas if R is well calibrated, only the selection weighted but not the unweighted TPR, FPR and AUC are equivalent to the respective measures in the training population, i.e. $TPR_0^W(r^*) = TPR_1(r^*)$, $FPR_0^W(r^*) = FPR_1$, and $AUC_0^W = AUC_1$.

Failure to transport

When $\pi_0(\mathbf{X}^*) \neq \pi_1(\mathbf{X}^*)$, i.e. the training a validation populations have different true disease probabilities, possibly because of differences in the distributions of unmeasured confounders, then $C_0^W = C_1$ only if $\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)dF_1(\mathbf{x}^*) = \int_{\mathbf{x}^*} \pi_1(\mathbf{x}^*)dF_1(\mathbf{x}^*)$. Otherwise

$$C_0^W = \frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)w(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*} R(\mathbf{x})w(\mathbf{x}^*)dF_0(\mathbf{x}^*)} = \frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}^*} R(\mathbf{x})dF_1(\mathbf{x}^*)} \neq C_1. \tag{4.23}$$

For the accuracy and discrimination measures, when $\pi_0(\mathbf{X}^*) \neq \pi_1(\mathbf{X}^*)$, we also have $TPR_0^W(r^*) \neq TPR_1(r^*)$, $FPR_0^W(r^*) \neq FPR_1(r^*)$, and $AUC_0^W \neq AUC_1$ if no additional assumptions have been made.

Table 4.1 summarizes the discussion of the reproducibility and transportability conditions in this section. When the true disease probabilities and the conditional distributions of the covariates given risk model predictors are the same, i.e, $\pi_1(\mathbf{X}^*) = \pi_0(\mathbf{X}^*)$ and $F_0(\mathbf{Z}|\mathbf{X}) = F_1(\mathbf{Z}|\mathbf{X})$, assessing if $C_0 = C_0^W = 1$ is the same as evaluating if the model is reproducible ($F_0(\mathbf{X}) = F_1(\mathbf{X})$). Given such assumptions, if only the selection weighted but not the unweighted calibration ratio is one, the model is transportable ($F_0(\mathbf{X}) \neq F_1(\mathbf{X})$). When the true disease probabilities are not the same, the model is neither reproducible nor transportable as the calibration ratio, regardless of being weighted or not, is not the same as the internal value C_1. If the true disease probabilities are the same but not the conditional distributions of the covariates given predictors in R, having $F_0(\mathbf{X}) = F_1(\mathbf{X})$ in addition does not help in the reproducibility of the model, while the model is not transportable either in this case.

In practice, when evaluating a model with external samples believing that their distributions of predictors are not the same as the one in the training population, we can conclude that the model is transportable if the selection weighted calibration ratio but not the unweighted one is equal to the internal calibration ratio, while not transportable if the selection weighted calibration ratio is not equal to the C_1 either.

Table 4.1: Unweighted (C_0) and weighted (C_0^W) calibration ratios and weighted AUC_0^W for risk model $R(\mathbf{X})$ under different settings. $\mathbf{X}^* = (\mathbf{X}, \mathbf{Z})$, $\pi_T(\mathbf{X}^*) = P_T(D = 1|\mathbf{X}^*)$, $\pi_T(\mathbf{X}) = \int_{\mathbf{z}} \pi_T(\mathbf{X}, \mathbf{z})dF_T(\mathbf{z}|\mathbf{X})$.

Relationship of true outcome probabilities and/or conditional distribution of covariates	Risk factors distributions	C_0	C_0^W	AUC_0^W
$\pi_1(\mathbf{X}^*) = \pi_0(\mathbf{X}^*)$ and $F_0(\mathbf{Z}\|\mathbf{X}) = F_1(\mathbf{Z}\|\mathbf{X})$	Reproducibility $F_0(\mathbf{X}) = F_1(\mathbf{X})$	$= C_1 = 1$	$= C_1 = 1$	$= AUC_1$
	Transportability $F_0(\mathbf{X}) \neq F_1(\mathbf{X})$	$\frac{\int_{\mathbf{x}^*} \pi_0(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_0(\mathbf{x})}$	$= C_1 = 1$	$= AUC_1$
$\pi_1(\mathbf{X}^*) \neq \pi_0(\mathbf{X}^*)$ or $\{\pi_1(\mathbf{X}^*) = \pi_0(\mathbf{X}^*)$ and $F_0(\mathbf{Z}\|\mathbf{X}) \neq F_1(\mathbf{Z}\|\mathbf{X})\}$	Reproducibility $F_0(\mathbf{X}) = F_1(\mathbf{X})$	$\neq C_1 = 1$	$\neq C_1 = 1$	$\neq AUC_1$
	Transportability $F_0(\mathbf{X}) \neq F_1(\mathbf{X})$	$\neq C_1 = 1$	$\neq C_1 = 1$	$\neq AUC_1$

Analytic example

We illustrate the performance of the unweighted and selection weighted calibration ratios using an analytic example with two binary risk factors X and Z, $\mathbf{X}^* = (X, Z)$. We let the true disease probabilities in the training and validation populations be the same, $\pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$, but the joint distributions of X and Z differ, i.e. $F_1(Z|X) \neq F_0(Z|X)$. The following Table 4.2 gives the numerical settings.

Table 4.2: Distributions of outcome D, model predictor X, and risk factor Z with resulting C_0 and C_0^W. $\pi_T(X, Z) = P_T(D = 1|X, Z)$.

Set T	X	$P_T(X)$	$P_T(Z = 1\|X)$	$\pi_T(X, Z = 0)$	$\pi_T(X, Z = 1)$	Measure
$T = 1$	1	0.9	0.8	0.05	0.1	$C_1 = 1$
(training)	0	0.1	0.2	0.01	0.05	
$T = 0$	1	0.2	0.2	0.05	0.1	$C_0 = 0.716$
(validation)	0	0.8	0.1	0.01	0.05	$C_0^W = 1$

We assume that the risk model R is only a function of X with values corresponding to the true probabilities in the training set, i.e.

$$\begin{aligned} R(X = 1) = P_1(D = 1|X = 1) &= \sum_z P_1(D = 1|X = 1, Z = z) P_1(Z = z|X = 1) \\ &= 0.05 \times 0.2 + 0.1 \times 0.8 = 0.09. \end{aligned}$$

Similarly, $R(X = 0) = 0.01 \times 0.8 + 0.05 \times 0.2 = 0.018$. Thus R is strongly calibrated in the training set and $C_1 = 1$.

In the validation set, the unweighted calibration ratio is

$$\begin{aligned} C_0 &= \frac{E_0(D)}{E_0[R(X)]} = \frac{\sum_x \sum_z P_0(D = 1|X = x, Z = z) P_0(Z = z|X = x) P_0(X = x)}{\sum_x \sum_z R(X = x) P_0(X = x, Z = z)} \\ &= \frac{0.0232}{0.0324} = 0.716. \end{aligned} \tag{4.24}$$

The selection weights $w(\mathbf{X}^*)$ given $\mathbf{X}^* = (X, Z)$ are computed from equation (4.11) and Table 4.2 using that $P_T(X = x, Z = z) = P_T(Z = z|X = x) P_T(X = x), T = 0, 1$. The values of $w(\mathbf{X}^*)$ for $(X, Z) = (1, 1), (1, 0), (0, 1), (0, 0)$ are $18, 9/8, 1/4, 1/9$, respectively, reflecting that validation set has much lower proportion of individuals with $(X, Z) = (1, 1)$ compared to the training data because the corresponding weight is the highest among the four combinations. Multiplying each summation term in the numerator and denominator in (4.24) by the weights yields

$$C_0^W = \frac{E_0[Dw(X^*)]}{E_0[R(X)w(X^*)]} = \frac{0.0828}{0.0828} = 1,$$

indicating that the model is transportable but not reproducible with such validation cohort as the selection weighted calibration ratio on the validation cohort is one while the unweighted ratio is away from one.

Estimating selection weighted performance measures

To estimate the selection weighted measures, we first calculate the selection weights w based on a model built on the combined cohort of training and validation sets using a binary cohort indicator as the response. We then plug in the computed selected weights into the empirical versions of the performance measures.

We estimate the selection weights based on $N_0 + N_1$ individuals in the pooled training and validation data via logistic regression. The model for the weights can include predictors $\mathbf{Z}$ in addition to the predictors $\mathbf{X}$ used in the risk model R:

$$P(T = 1|\mathbf{X}^*, \boldsymbol{\gamma}) = \exp(\gamma_0 + \boldsymbol{\gamma}_1{}'\mathbf{X} + \boldsymbol{\gamma}_2{}'\mathbf{Z})/\{1 + \exp(\gamma_0 + \boldsymbol{\gamma}_1{}'\mathbf{X} + \boldsymbol{\gamma}_2{}'\mathbf{Z})\}, \tag{4.25}$$

where $\mathbf{X}^* = (\mathbf{X}, \mathbf{Z})$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are the log-odds ratios for risk factors $\mathbf{X}$ and $\mathbf{Z}$ respectively and $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)'$. We then compute the weights for participants in the validation cohort

$$\widehat{w}(\mathbf{x}_i^*) = w(\mathbf{x}_i^*, \widehat{\boldsymbol{\gamma}}) = \frac{\widehat{P}(T = 1|\mathbf{x}_i^*)N_0}{\widehat{P}(T = 0|\mathbf{x}_i^*)N_1} = \exp(\widehat{\gamma}_0 + \widehat{\boldsymbol{\gamma}}_1'\mathbf{x}_i + \widehat{\boldsymbol{\gamma}}_2'\mathbf{z}_i)\frac{N_0}{N_1}, i = 1, \ldots, N_0, \tag{4.26}$$

where $\widehat{\boldsymbol{\gamma}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\gamma}$.

The estimate of the selection weighted calibration ratio on validation cohort is

$$\hat{C}_0^W = \frac{\sum_{i=1}^{N_0} D_i \widehat{w}(\mathbf{x}_i^*)}{\sum_{i=1}^{N_0} R(\mathbf{x}_i)\widehat{w}(\mathbf{x}_i^*)}. \tag{4.27}$$

The estimates of the selection weighted TPR and FPR are

$$\widehat{TPR}_0^W(r^*) = \frac{\sum_{i=1}^{N_0} \widehat{w}(\mathbf{x}_i^*) I(R(\mathbf{x}_i) \geq r^*) D_i}{\sum_{i=1}^{N_0} D_i \widehat{w}(\mathbf{x}_i^*)} \tag{4.28}$$

$$FPR_0^W(r^*) = \frac{\sum_{i=1}^{N_0} \widehat{w}(\mathbf{x}_i^*) I(R(\mathbf{x}_i) \geq r^*)(1 - D_i)}{\sum_{i=1}^{N_0} (1 - D_i)\widehat{w}(\mathbf{x}_i^*)}. \tag{4.29}$$

The estimate of the selection weighted AUC accounting for ties is

$$\widehat{AUC}_0^W = \frac{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i(1 - D_j)\hat{w}(\mathbf{x}_i^*)\hat{w}(\mathbf{x}_j^*)[I(R(\mathbf{x}_i) > R(\mathbf{x}_j)) + 0.5I(R(\mathbf{x}_i) = R(\mathbf{x}_j))]}{\sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i(1 - D_j)\hat{w}(\mathbf{x}_i^*)\hat{w}(\mathbf{x}_j^*)}. \tag{4.30}$$

Note that the constant N_0/N_1 in (4.26) cancels out in all estimates as it appears in both the numerator and denominator of (4.27) - (4.30). The estimates (4.27) - (4.30) are consistent for the respective population measures and asymptotically normally distributed, which proof follows similarly as showing the large sample properties in Section 3.2. Here, though we have the proposed selection weights in a form different to the example form of weights we check in the Section 3.2, they are also from logistic regression. Therefore, the arguments in Section 3.2 apply to here as well.

Accommodate verification bias

In addition to the previously discussed selection bias arising from different distributions of model risk factors, we discuss the impact of differences in disease verification procedures between the training and validation populations, which affects the observed probabilities of disease status D.

We denote $V = 1$ if the verification test to diagnose disease has been performed and $V = 0$ otherwise. The true probability of disease is

$$\pi_T(\mathbf{X}^*) = P_T(D = 1|\mathbf{X}^*) = P_T(D = 1, V = 1|\mathbf{X}^*) + P_T(D = 1, V = 0|\mathbf{X}^*), T = 0, 1. \tag{4.31}$$

However, we only observe the definitive disease status D when the verification test has been performed but not for the others, i.e. we only observe $P(D = 1, V = 1|\mathbf{X}^*, T) = P(D = 1|\mathbf{X}^*, T, V = 1)P(V = 1|\mathbf{X}^*, T)$ in practice.

First, we assume $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$, such that the only difference between the training and validation cohorts is the verification process. We combine disease verification bias adjustment with adjustment for selection bias discussed in the previous section. If R is developed in the training data using only verified disease as the outcome, then R is actually a model for $P_1(D = 1, V = 1|\mathbf{X}^*) = P_1(DV = 1|\mathbf{X}^*)$ but not for $\pi_1(\mathbf{X}^*) = P_1(D = 1|\mathbf{X}^*)$. Here, $DV = 1$ when event $\{D = 1, V = 1\}$ occurs, while $DV = 0$ corresponds to the events $\{V = 0\} \cup \{D = 0, V = 1\}$, i.e. being not verified or having no disease and being verified. The $\cup$ refers to the union of the events. We use a tilde in this section to distinguish the measures from the ones presented in the earlier sections. We assume the model is well calibrated in the training population, i.e.

$$\widetilde{C}_1 = \frac{E_1(DV)}{E_1[R(\mathbf{X})]} = \frac{\int_{\mathbf{x}^*} P_1(D = 1|\mathbf{x}^*, V = 1)P_1(V = 1|\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_1(\mathbf{x})} = 1. \tag{4.32}$$

To assess the impact of differences in verification process between the training and validation data on calibration, we define verification weights as

$$v(\mathbf{X}^*) = \frac{P_1(V = 1|\mathbf{X}^*)}{P_0(V = 1|\mathbf{X}^*)} \text{ and } \overline{v}(\mathbf{X}^*) = \frac{P_1(V = 0|\mathbf{X}^*)}{P_0(V = 0|\mathbf{X}^*)}, \tag{4.33}$$

The weight $\overline{v}$ is only used in the computation of performance measures that rely on verified individuals without detected disease or unverified individuals, i.e. $DV = 0$, such as in the calculation of FPR and AUC.

The *verification weighted* calibration ratio in the validation population is defined as

$$\widetilde{C}_0^V = \frac{E_0[DVv(\mathbf{X}^*)]}{E_0[R(\mathbf{X})]} = \frac{\int_{\mathbf{x}^*} P_0(D=1|\mathbf{x}^*, V=1)P_0(V=1|\mathbf{x}^*)v(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_0(\mathbf{x})}. \tag{4.34}$$

If $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$, $P_0(D=1|\mathbf{X}^*, V=1) = P_1(D=1|\mathbf{X}^*, V=1)$ or $P_0(D=1|\mathbf{X}^*) = \pi_0(\mathbf{X}^*) = \pi_1(\mathbf{X}^*)$ if D is conditionally independent of V given $\mathbf{X}^*$, i.e. $P_T(D=1|\mathbf{X}^*, V) = P_T(D=1|\mathbf{X}^*)$, we have

$$\begin{aligned}\widetilde{C}_0^V &= \frac{E_0[DVv(\mathbf{X}^*)]}{E_0[R(\mathbf{X})]} = \frac{\int_{\mathbf{x}^*} P_0(D=1|\mathbf{x}^*, V=1)P_0(V=1|\mathbf{x}^*)v(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_0(\mathbf{x})} \\ &= \frac{\int_{\mathbf{x}^*} P_1(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}} R(\mathbf{x})dF_1(\mathbf{x})} = \widetilde{C}_1 = 1, \end{aligned} \tag{4.35}$$

given that the model for v is correctly specified. Unlike the case adjusting for selection bias that we weight both predictions and observations, we weight the observations but not the predictions here. The formula for the verification weighted TPR is

$$\begin{aligned}\widetilde{TPR}_0^V(r^*) &= \frac{E_0\{v(\mathbf{X}^*)I(R(\mathbf{X}) \geq r^*)DV\}}{E_0\{DVv(\mathbf{X}^*)\}} \\ &= \frac{\int_{\mathbf{x}^*} I(R(\mathbf{x}) \geq r^*)P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*} P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_0(\mathbf{x}^*)}. \end{aligned} \tag{4.36}$$

Therefore, if $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$ and $P_0(D=1|\mathbf{X}^*, V=1) = P_1(D=1|\mathbf{X}^*, V=1)$, we can show

$$\begin{aligned}\widetilde{TPR}_0^V(r^*) &= \frac{\int_{\mathbf{x}^*} I(R(\mathbf{x}) \geq r^*)P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*} P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_0(\mathbf{x}^*)} \\ &= \frac{\int_{\mathbf{x}^*} I(R(\mathbf{x}) \geq r^*)P_1(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_1(\mathbf{x}^*)}{\int_{\mathbf{x}^*} P_1(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)dF_1(\mathbf{x}^*)} \\ &= \widetilde{TPR}_1(r^*). \end{aligned} \tag{4.37}$$

To obtain the verification adjusted FPR, we first note that by Bayes theorem,

$$F_T(\mathbf{X}^*|DV=0) = \frac{\{P_T(V=0|\mathbf{X}^*) + P_T(D=0|\mathbf{X}^*, V=1)P_T(V=1|\mathbf{X}^*)\}F_T(\mathbf{X}^*)}{1 - E_T(DV)}, \tag{4.38}$$

where we use $P(DV=0) = 1 - P(DV=1) = 1 - E(DV)$. Thus, we have

$$\widetilde{FPR}_T(r^*) = P_T(R(\mathbf{X}) \geq r^*|DV=0) = \frac{E_T\left[I(R(\mathbf{X}) \geq r^*)\{(1-V) + (1-D)V\}\right]}{E_T\{(1-V) + (1-D)V\}}. \tag{4.39}$$

The first term in the numerator of (4.39) captures the contribution of the unverified individuals to the FPR, and the second term captures the contribution of those who were verified and found to be not diseased, i.e. $V=1$ and $D=0$. Thus the verification weighted version is

$$\widetilde{FPR}_0^V(r^*) = \frac{E_0\big[I(R(\mathbf{X}) \geq r^*)\{v(\mathbf{X}^*)(1-D)V + \overline{v}(\mathbf{X}^*)(1-V)\}\big]}{E_0[v(\mathbf{X}^*)(1-D)V + \overline{v}(\mathbf{X}^*)(1-V)]}. \qquad (4.40)$$

We can write it as

$$\begin{aligned} & \widetilde{FPR}_0^V(r^*) \\ = & \frac{\int_{\mathbf{x}^*} I(R(\mathbf{x}) \geq r^*)[P_0(V=0|\mathbf{x}^*)\overline{v}(\mathbf{x}^*) + P_0(D=1|\mathbf{x}^*, V=1)P_0(V=1|\mathbf{x}^*)v(\mathbf{x}^*)]dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*}[P_0(V=0|\mathbf{x}^*)\overline{v}(\mathbf{x}^*) + P_0(D=1|\mathbf{x}^*, V=1)P_0(V=1|\mathbf{x}^*)v(\mathbf{x}^*)]dF_0(\mathbf{x}^*)} \\ = & \frac{\int_{\mathbf{x}^*} I(R(\mathbf{x}) \geq r^*)[P_1(V=0|\mathbf{x}^*) + P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)]dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}^*}[P_1(V=0|\mathbf{x}^*) + P_0(D=1|\mathbf{x}^*, V=1)P_1(V=1|\mathbf{x}^*)]dF_0(\mathbf{x}^*)}. \qquad (4.41) \end{aligned}$$

When $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$ and $P_0(D=0|\mathbf{X}^*, V=1) = P_1(D=0|\mathbf{X}^*, V=1)$, we have $\widetilde{FPR}_0^V(r^*) = \widetilde{FPR}_1(r^*)$.

Using Bayes theorem and recalling that the risk model is based only on $\mathbf{X}$ but the true disease probability depends on $\mathbf{X}^*$, we have

$$\begin{aligned} \widetilde{AUC}_T &= \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\}dF_T(\mathbf{u}^*|DV=1)dF_T(\mathbf{y}^*|DV=0) \\ &= \frac{S_T^1 + S_T^2}{E_T(DV)E_T(1-DV)}, T = 0, 1, \qquad (4.42) \end{aligned}$$

where

$$S_T^1 = \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\}P_T(D=1|\mathbf{u}^*, V=1)P_T(V=1|\mathbf{u}^*)P_T(V=0|\mathbf{y}^*)dF_T(\mathbf{u}^*)dF_T(\mathbf{y}^*), \qquad (4.43)$$

and

$$\begin{aligned} S_T^2 = \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} & I\{R(\mathbf{u}) > R(\mathbf{y})\}P_T(D=1|\mathbf{u}^*, V=1)P_T(V=1|\mathbf{u}^*) \\ & P_T(D=0|\mathbf{y}^*, V=1)P_T(V=1|\mathbf{y}^*)dF_T(\mathbf{u}^*)dF_T(\mathbf{y}^*). \qquad (4.44) \end{aligned}$$

Similar to the FPR, the S_T^1 captures the contribution of the unverified individuals to the AUC, and S_T^2 is the contribution of those verified and found to be not diseased. Here, to obtain S_T^1 and S_T^2, we first write the integral in (4.42) as

$$\begin{aligned} & \widetilde{AUC}_T = \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\}dF_T(\mathbf{u}^*|DV=1)dF_T(\mathbf{y}^*|DV=0) \\ = & \frac{1}{P_T(DV=1)}\frac{1}{P_T(DV=0)}\int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\}dF_T(\mathbf{u}^*, DV=1)dF_T(\mathbf{y}^*, DV=0) \\ = & \frac{1}{E_T(DV)E_T(1-DV)}\int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\}dF_T(\mathbf{u}^*, DV=1)dF_T(\mathbf{y}^*, DV=0). \end{aligned}$$

Because $\{DV = 0\}$ consists of events $\{V = 0\}$ and $\{D = 0, V = 1\}$, we have

$$\begin{aligned} dF_T(\mathbf{u}^*, DV = 1) &= \frac{P_T(DV = 1, \mathbf{u}^*)}{P_T(\mathbf{u}^*)} P_T(\mathbf{u}^*) \\ &= P_T(DV = 1|\mathbf{u}^*) dF_T(\mathbf{u}^*) \\ &= P_T(D = 1|\mathbf{u}^*, V = 1) P(V = 1|\mathbf{u}^*) dF_T(\mathbf{u}^*). \end{aligned} \tag{4.45}$$

For $dF_T(\mathbf{y}^*, DV = 0)$, we have

$$\begin{aligned} dF_T(\mathbf{y}^*, DV = 0) &= P_T(DV = 0|\mathbf{y}^*) dF_T(\mathbf{y}^*) \\ &= [P_T(V = 0|\mathbf{y}^*) + P_T(D = 0|\mathbf{y}^*, V = 1) P_T(V = 1|\mathbf{y}^*)] dF_T(\mathbf{y}^*). \end{aligned}$$

Then, we can give the expressions of S_T^1 and S_T^2 as (4.43) and (4.44), respectively.

Under the assumptions that $P_0(D = 1|\mathbf{X}^*, V = 1) = P_1(D = 1|\mathbf{X}^*, V = 1)$ and $F_0(\mathbf{x}^*) = F_1(\mathbf{x}^*)$, $E_0[DVv(\mathbf{X}^*)] = \int P_1(D = 1|\mathbf{x}^*, V = 1) P_1(V = 1|\mathbf{x}) dF_1(\mathbf{x}) = E_1(DV)$. We then have

$$S_0^{V1} = \int_{\mathbf{u}^*} \int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\} v(\mathbf{u}^*) \overline{v}(\mathbf{y}^*) P_0(D = 1|\mathbf{u}^*, V = 1) P_0(V = 1|\mathbf{u}^*)$$
$$P_T(V = 0|\mathbf{y}^*) dF_0(\mathbf{u}^*) dF_0(\mathbf{y}^*) = S_1^1, \tag{4.46}$$

$$S_0^{V2} = \int_{\mathbf{u}^*} \int_{\mathbf{y}^*} I\{R(\mathbf{u}) > R(\mathbf{y})\} v(\mathbf{u}^*) v(\mathbf{y}^*) P_0(D = 1|\mathbf{u}^*, V = 1) P_0(V = 1|\mathbf{u}^*)$$
$$P_0(D = 0|\mathbf{y}^*, V = 1) P_0(V = 1|\mathbf{y}^*) dF_0(\mathbf{u}^*) dF_0(\mathbf{y}^*) = S_1^2. \tag{4.47}$$

The verification weighted AUC is defined as:

$$\widetilde{AUC}_0^V = \frac{S_0^{V1} + S_0^{V2}}{E_0(DVv(\mathbf{X}^*)) E_0\{\overline{v}(\mathbf{X}^*)(1 - V) + v(\mathbf{X}^*)(1 - D)V\}}, \tag{4.48}$$

which is the same as the AUC in the training set ($\widetilde{AUC}_1$) given the above assumptions hold.

Failure to validate in the presence of verification differences

$\widetilde{C}_0^V$ in (4.35) is not equal to one, if $P_0(D = 1|\mathbf{x}^*, V = 1) \neq P_1(D = 1|\mathbf{x}^*, V = 1)$ or $P_0(D = 1|\mathbf{X}^*) = \pi_0(\mathbf{X}^*) \neq \pi_1(\mathbf{X}^*)$ when D is conditionally independent of V given $\mathbf{X}^*$. All above assume that the weights v are correctly modeled and $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$. $\widetilde{C}_0^V$ can also not equal to one if either of these two conditions does not hold, i.e. the weights v are not correctly modeled or $F_0(\mathbf{X}^*) \neq F_1(\mathbf{X}^*)$.

Performance assessment only in the verified population

We can also restrict the model assessment to only the verified outcomes depending on the target population and the investigation purpose. For example, several models predict the probability of prostate cancer in individuals who are verified via prostate biopsy [Ankerst et al., 2018]. In such application, $R^{V=1}(\mathbf{X})$ estimates $P(D = 1|\mathbf{X}, V = 1)$ and the performance assessment is also restricted to individuals with $V = 1$. In the diagnostic testing literature, this is known as *complete case analysis* [Alonzo and Pepe, 2005]. The unweighted calibration ratio restricted to the verified participants in the validation set is

$$C_0(R^{V=1}) = \frac{E_0\{DV\}}{E_0\{R^{V=1}(\mathbf{X})V\}} = \frac{\int_{\mathbf{x}^*} P_0(D = 1|\mathbf{x}^*, V = 1)P_0(V = 1|\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R^{V=1}(\mathbf{x})P_0(V = 1|\mathbf{x}^*)dF_0(\mathbf{x}^*)} \tag{4.49}$$

and the corresponding verification weighted measure is

$$C_0^V(R^{V=1}) = \frac{E_0\{DVv(\mathbf{X}^*)\}}{E_0\{R^{V=1}(\mathbf{X})V\}} = \frac{\int_{\mathbf{x}^*} P_0(D = 1|\mathbf{x}^*, V = 1)P_0(V = 1|\mathbf{x}^*)v(\mathbf{x}^*)dF_0(\mathbf{x}^*)}{\int_{\mathbf{x}} R^{V=1}(\mathbf{x})P_0(V = 1|\mathbf{x}^*)v(\mathbf{x}^*)dF_0(\mathbf{x}^*)}. \tag{4.50}$$

Similar to C_0^V that is computed in the overall population, when $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$ and $P_0(D = 1|\mathbf{x}^*, V = 1) = P_1(D = 1|\mathbf{x}^*, V = 1)$, then $C_0^V(R^{V=1}) = C_1(R^{V=1})$.

Among those who have been verified including diseased and non-diseased but verified individuals, the unweighted AUC is computed as

$$AUC_0(R^{V=1}) = \frac{1}{E_0(DV)E_0(1-DV)} \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R^{V=1}(\mathbf{u}) > R^{V=1}(\mathbf{y})\}$$
$$P_0(D = 1|\mathbf{u}^*, V = 1)P_0(V = 1|\mathbf{u}^*)P_0(D = 0|\mathbf{y}^*, V = 1)P_0(V = 1|\mathbf{y}^*)dF_0(\mathbf{u}^*)dF_0(\mathbf{y}^*). \tag{4.51}$$

The corresponding verification weighted quantity is

$$AUC_0^V(R^{V=1}) = \frac{1}{E_0(DV)E_0(1-DV)} \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} I\{R^{V=1}(\mathbf{u}) > R^{V=1}(\mathbf{y})\}$$
$$P_0(D = 1|\mathbf{u}^*, V = 1)P_0(V = 1|\mathbf{u}^*)v(\mathbf{u}^*)P_0(D = 0|\mathbf{y}^*, V = 1)P_0(V = 1|\mathbf{y}^*)v(\mathbf{y}^*)dF_0(\mathbf{u}^*)dF_0(\mathbf{y}^*). \tag{4.52}$$

Here only v but not $\bar{v}$ is used in $AUC_0^V(R^{V=1})$. The same as the situation for the calibration measure, $AUC_0^V(R^{V=1}) = AUC_1(R^{V=1})$ when $F_0(\mathbf{X}^*) = F_1(\mathbf{X}^*)$ and $P_0(D = 1|\mathbf{x}^*, V = 1) = P_1(D = 1|\mathbf{x}^*, V = 1)$.

Estimating verification weighted performance measures

The verification weights are estimated by computing $P_T(V = 1|\mathbf{X}^*)$ or $P_T(V(\tau) = 1|\mathbf{X}^*), T = 0, 1$, for models that predict over a pre-specified period τ, separately in the training and verification sets using logistic regression models

$$\widehat{P_T}(V = 1|\mathbf{X}^*) = \exp(\eta_{T0} + +\boldsymbol{\eta}'_{T1}\mathbf{X} + \boldsymbol{\eta}'_{T2}\mathbf{Z})/\{1 + \exp(\eta_{T0} + +\boldsymbol{\eta}'_{T1}\mathbf{X} + \boldsymbol{\eta}'_{T2}\mathbf{Z})\}, \tag{4.53}$$

and then taking the ratio. Alternatively one could fit a survival model $S(t, \mathbf{X}^*)$ to the time to disease verification and compute $\widehat{P}_T(V(\tau) = 1|\mathbf{X}^*) = 1 - \widehat{S}(\tau, \mathbf{X}^*), T = 0, 1$.The final weights are

$$\widehat{v}(\mathbf{X}^*) = \frac{\widehat{P}_1(V = 1|\mathbf{X}^*)}{\widehat{P}_0(V = 1|\mathbf{X}^*)} \text{ and } \widehat{\bar{v}}(\mathbf{X}^*) = \frac{\widehat{P}_1(V = 0|\mathbf{X}^*)}{\widehat{P}_0(V = 0|\mathbf{X}^*)}. \tag{4.54}$$

We replace the expressions in (4.34) and (4.48) by their empirical quantities to obtain estimates of the verification weighted performance measures. For notation simplicity, we ignore the tilde in the notation for verification weighted measures calculated on the validation cohort from now on. We estimate them as:

$$\widehat{C}_0^V = \frac{\sum_{i=1}^{N_0} \widehat{v}(\mathbf{x}_i^*) D_i V_i}{\sum_{i=1}^{N_0} R(\mathbf{x}_i)}, \tag{4.55}$$

$$\widehat{TPR}_0^V(r^*) = \frac{\sum\limits_{i=1}^{N_0} \widehat{v}(\mathbf{x}_i^*) I(R(\mathbf{x}_i) \geq r^*) D_i V_i}{\sum\limits_{i=1}^{N_0} \widehat{v}(\mathbf{x}_i^*) D_i V_i}, \tag{4.56}$$

$$\widehat{FPR}_0^V(r^*) = \frac{\sum\limits_{i=1}^{N_0} I(R(\mathbf{x}_i^*) \geq r^*)\{\widehat{v}(\mathbf{x}_i^*)(1 - D_i)V_i + \widehat{\bar{v}}(\mathbf{x}_i^*)(1 - V_i)\}}{\sum\limits_{i=1}^{N_0} \{\widehat{\bar{v}}(\mathbf{x}_i^*)(1 - V_i) + \widehat{v}(\mathbf{x}_i^*)(1 - D_i)V_i\}}, \tag{4.57}$$

and

$$\widehat{AUC}_0^V = \frac{\widehat{S}_0^{V1} + \widehat{S}_0^{V2}}{\{\sum\limits_{i=1}^{N_0} D_i V_i \widehat{v}(\mathbf{x}_i^*)\} \sum\limits_{i=1}^{N_0} \{\widehat{\bar{v}}(\mathbf{x}_i^*)(1 - V_i) + \widehat{v}(\mathbf{x}_i^*)(1 - D_i)V_i\}}, \tag{4.58}$$

where

$$\widehat{S}_0^{V1} = \sum_{i=1}^{N_0} \sum_{j=1}^{N_0} D_i V_i (1 - V_j) \widehat{v}(\mathbf{x}_i^*) \widehat{\bar{v}}(\mathbf{x}_j^*) [I(R(\mathbf{x}_i) > R(\mathbf{x}_j)) + 0.5 I(R(\mathbf{x}_i) = R(\mathbf{x}_j))], \tag{4.59}$$

and the verification weighted version of the second term is

$$\widehat{S}_0^{V2} = \sum_{i=1}^{N_0} \sum_{j=1}^{N_0} D_i V_i (1 - D_j) V_j \widehat{v}(\mathbf{x}_i^*) \widehat{v}(\mathbf{x}_j^*) [I(R(\mathbf{x}_i) > R(\mathbf{x}_j)) + 0.5 I(R(\mathbf{x}_i) = R(\mathbf{x}_j))]. \tag{4.60}$$

The consistency and asymptotic normality hold for these estimators with similar arguments as shown in Section 3.2.

Joint weighting with selection and verification weights

In a final step, we combine the selection and verification weights for all the measures, adjusting for both selection and verification bias. The combined weighted measures are

$$C_0^{WV} = \frac{E_0\{DVv(\mathbf{X}^*)w(\mathbf{X}^*)\}}{E_0\{R(\mathbf{X})w(\mathbf{X}^*)\}}, \tag{4.61}$$

$$TPR_0^{WV}(r^*) = \frac{E_0\{w(\mathbf{X}^*)v(\mathbf{X}^*)I(R(\mathbf{X}) \geq r^*)DV\}}{E_0\{DVw(\mathbf{X}^*)v(\mathbf{X}^*)\}}, \tag{4.62}$$

$$\begin{aligned} FPR_0^{WV}(r^*) = \frac{1}{E_0\{w(\mathbf{X}^*)\overline{v}(\mathbf{X}^*)(1-DV)\}} \big[& E_0\{I(R(\mathbf{X}) \geq r^*)w(\mathbf{X}^*)\overline{v}(\mathbf{X}^*)(1-V)\} \\ & + E_0\{I(R(\mathbf{X}) \geq r^*)(1-D)Vw(\mathbf{X}^*)v(\mathbf{X}^*)\}\big], \end{aligned} \tag{4.63}$$

and

$$AUC_0^{WV} = \frac{S_0^{WV1} + S_0^{WV2}}{E_0\{DVw(\mathbf{X}^*)v(\mathbf{X}^*)\}E_0\{w(\mathbf{X}^*)\overline{v}(\mathbf{X}^*)(1-DV)\}}, \tag{4.64}$$

where

$$\begin{aligned} S_0^{WV1} = \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} & I\{R(\mathbf{u}) > R(\mathbf{y})\}P_0(D=1|\mathbf{u}^*, V=1)P_0(V=1|\mathbf{u}^*)P_0(V=0|\mathbf{y}^*) \\ & w(\mathbf{u}^*)w(\mathbf{y}^*)v(\mathbf{u}^*)\overline{v}(\mathbf{y}^*)dF_0(\mathbf{u}^*)dF_0(\mathbf{y}^*), \end{aligned} \tag{4.65}$$

and

$$\begin{aligned} S_0^{WV2} = \int_{\mathbf{u}^*}\int_{\mathbf{y}^*} & I\{R(\mathbf{u}) > R(\mathbf{y})\}P_0(D=1|\mathbf{u}^*, V=1)P_0(V=1|\mathbf{u}^*)P_0(D=0|\mathbf{y}^*, V=1) \\ & P_0(V=1|\mathbf{y}^*)w(\mathbf{u}^*)w(\mathbf{y}^*)v(\mathbf{u}^*)v(\mathbf{y}^*)dF_0(\mathbf{u}^*)dF_0(\mathbf{y}^*). \end{aligned} \tag{4.66}$$

The estimates of the jointly weighted performance measures are

$$\widehat{C_0}^{WV} = \frac{\sum_{i=1}^{N_0} \widehat{w}(\mathbf{x}_i^*)\widehat{v}(\mathbf{x}_i^*)D_iV_i}{\sum_{i=1}^{N_0} R(\mathbf{x}_i)\widehat{w}(\mathbf{x}_i^*)}, \tag{4.67}$$

$$\widehat{TPR_0}^{WV}(r^*) = \frac{\sum\limits_{i=1}^{N_0} \widehat{v}(\mathbf{x}_i^*)\widehat{w}(\mathbf{x}_i^*)I(R(\mathbf{x}_i) \geq r^*)D_iV_i}{\sum\limits_{i=1}^{N_0} \widehat{w}(\mathbf{x}_i^*)\widehat{v}(\mathbf{x}_i^*)D_iV_i}, \tag{4.68}$$

$$\widehat{FPR}_0^{WV}(r^*) = \frac{\sum_{i=1}^{N_0} I(R(\mathbf{x}_i^*) \geq r^*)\widehat{w}(\mathbf{x}_i^*)\{\widehat{v}(\mathbf{x}_i^*)(1-D_i)V_i + \widehat{\overline{v}}(\mathbf{x}_i^*)(1-V_i)\}}{\sum_{i=1}^{N_0} \widehat{w}(\mathbf{x}_i^*)\{\widehat{\overline{v}}(\mathbf{x}_i^*)(1-V_i) + \widehat{v}(\mathbf{x}_i^*)(1-D_i)V_i\}}, \tag{4.69}$$

and

$$\widehat{AUC}_0^{WV} = \frac{\widehat{S}_0^{WV1} + \hat{S}_0^{WV2}}{\{\sum_{i=1}^{N_0} D_i V_i \widehat{w}(\mathbf{x}_i^*)\widehat{v}(\mathbf{x}_i^*)\} \sum_{i=1}^{N_0}\{\widehat{w}(\mathbf{x}_i^*)\widehat{\overline{v}}(\mathbf{x}_i^*)(1-V_i) + \widehat{w}(\mathbf{x}_i^*)\widehat{v}(\mathbf{x}_i^*)(1-D_i)V_i\}}, \tag{4.70}$$

where

$$\widehat{S}_0^{WV1} = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i V_i (1-V_j)\widehat{w}(\mathbf{x}_i^*)\widehat{w}(\mathbf{x}_j^*)\widehat{v}(\mathbf{x}_i^*)\widehat{\overline{v}}(\mathbf{x}_j^*)[I(R(\mathbf{x}_i) > R(\mathbf{x}_j)) + 0.5 I(R(\mathbf{x}_i) = R(\mathbf{x}_j))], \tag{4.71}$$

and the weighted version of the second term is

$$\widehat{S}_0^{WV2} = \sum_{i=1}^{N_0}\sum_{j=1}^{N_0} D_i V_i (1-DV_j)\widehat{w}(\mathbf{x}_i^*)\widehat{w}(\mathbf{x}_j^*)\widehat{v}(\mathbf{x}_i^*)\widehat{v}(\mathbf{x}_j^*)[I(R(\mathbf{x}_i) > R(\mathbf{x}_j)) + 0.5 I(R(\mathbf{x}_i) = R(\mathbf{x}_j))]. \tag{4.72}$$

Consistency and asymptotic normality of the jointly weighted estimates follow as before.

4.3 Simulation study

We conduct a simulation study to explore the proposed weighted methods to adjust the selection bias or verification bias or both. We first generate training cohorts with $N_1 = 30000$ individuals. We then build the risk model $R(X)$ (later with $R(X, Z_1)$ as sensitivity analysis) with the training data and validate it under different validation scenarios. To compute the performance measures, for each validation scenario, we simulate $B = 500$ validation samples with each containing $N_0 = 20000$ individuals.

Configuration

We first describe the distributions used to generate the training data, followed by the risk model configuration. Then, we give the distributions used in simulating the validation cohorts and the method used to add on verification for both training and validation cohorts. We end this section with the settings of the models for the selection and verification weights.

Training data generation

The training cohort consists of $N_1 = 30000$ individuals with continuous risk factors X, Z_1, Z_2, where X is used in the risk prediction model $R(X)$. The vector of risk factors (X, Z_1, Z_2) is sampled from a multivariate normal (MVN) distribution with mean $(0, 0, 0)$ and covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.5 \\ 0.1 & 0.5 & 1 \end{pmatrix}. \tag{4.73}$$

We let $c = 50$ as the maximum follow-up time and $S_{max} = 0.98$, i.e. 98% of individuals survive until administrative censoring at time c. We generate the survival time Y^* given the predictors X and Z_1 of each individual by an exponential probability distribution with the hazard rate of

$$\lambda = \lambda_0 \exp(\beta_1 X + \beta_2 Z_1), \tag{4.74}$$

where $\lambda_0 = -\log(S_{max})/c \approx 4.0 \times 10^{-4}$ is the baseline hazard rate and $\beta_1 = \log(2)$ and $\beta_2 = \log(1.2)$. We allow individuals to exit the cohort at any time Y_D before c, e.g. death due to causes other than event of interest. Y_D is sampled from another exponential distribution with hazard rate

$$\lambda = -\log(S^c_{max})/Y_{max} \approx 2.2 \times 10^{-3}, \tag{4.75}$$

where $S^c_{max} = 0.9$ is the proportion of individuals who do not lost during follow-up. Incorporating administrative censoring at $c = 50$ and loss during follow-up, the observed event time is given by $Y = \min(Y^*, Y_D, c = 50)$ and the event indicator is $D = I(Y = Y^*)$.

Risk model estimation

We obtain the risk model R by fitting a Cox regression model as in (4.78) to the training cohort, but only using the risk factor X as the predictor, i.e. omitting Z_1 that also impacts disease risk in the population. In sensitivity analysis, we model the risk via $R(X, Z_1)$, which model is correctly specified and includes the main effects of X and Z_1.

Validation data generation

We generate validation cohorts of size $N_0 = 20000$ using the same data generation mechanism as described above for the training data but with several different distributions of (X, Z_1, Z_2). We study three different scenarios for validation samples: Scenario $S1$, where $(X, Z_1, Z_2) \sim MVN((0, 0, 0), \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ given in (4.73); Scenario $S2$, where $(X, Z_1, Z_2) \sim MVN((0, 0.5, 0.5), \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ given in (4.73); Scenario $S3$ with $(X, Z_1, Z_2) \sim MVN((0, 0, 0), \mathbf{\Sigma}_2)$ where

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 1 & 0.6 & -0.4 \\ 0.6 & 1 & -0.2 \\ -0.4 & -0.2 & 1 \end{pmatrix}. \tag{4.76}$$

Adding verification to training and validation data

To add on disease verification to the training ($T = 1$) and validation ($T = 0$) cohorts, we generate the verification status from binomial distribution: $V_i \sim Binom(1, p_T^{V_i})$ for each individual $i, i = 1, \dots, N_T$, which binomial distribution has the probability

$$p_T^{V_i} = P_T(V_i = 1|X_i, Z_{1i}) = \frac{\exp(\eta_{T0} + \eta_{T1}X_i + \eta_{T2}Z_{1i})}{1 + \exp(\eta_{T0} + \eta_{T1}X_i + \eta_{T2}Z_{1i})}, T = 0, 1. \tag{4.77}$$

We estimate the model $R(X)$ based on all individuals in the training cohort, $T = 1$, using the observed event times and the observed event indicator DV. The model R is then evaluated on all individuals in the validation cohort ($T = 1$) based on the event indicator DV.

Let $\boldsymbol{\eta}_T = (\eta_{T0}, \eta_{T1}, \eta_{T2}), T = 0, 1.$ we first assume the same verification mechanism is used in both training and validation data, i.e. $\boldsymbol{\eta}_0 = \boldsymbol{\eta}_1 = (-1.4, 1.3, -0.1)$ (Scenario $V1$). Second, we vary only the intercept in (4.77) with different sign, but all the other parameters are the same, i.e. $\boldsymbol{\eta}_0 = (-1.4, 1.3, -0.1)$ and $\boldsymbol{\eta}_1 = (1.4, 1.3, -0.1)$ (Scenario $V2$). For Scenario $V3$, all parameters used in (4.77) differ with $\boldsymbol{\eta}_0 = (-1.4, 1.3, -0.1)$ and $\boldsymbol{\eta}_1 = (1.4, -1.3, 0.1)$.

Weights and performance measures

We also explore different models used for computing the weights. The selection weights w_1 and verification weights v_1 come from logistic regressions including the main effects of X and Z_1, their interaction term, and their quadratic effects. The w_2 and v_2 come from logistic regressions including (X, Z_1, Z_2) as the main effects, their interaction terms, and their quadratic effects. The w_3 and v_3 use logistic regressions with (X, Z_2) as the main effects, their interaction term, and their quadratic effects, where the important variable, Z_1, is omitted in this setting.

Simulation results

In the simulation, we generate three different training cohorts each with setting $S1, V1$, $S1, V2$, and $S1, V3$, respectively. In contrast, for each scenario shown in Table 4.3, we generate the validation cohort 500 times and estimate the unweighted and weighted performance measures on the validation cohort each time. The mean values and standard deviations, given in the row below the mean values, of the estimated weighted and unweighted calibration ratios and AUCs with various scenarios and different weighting strategies are summarized in Table 4.3. The risk model $R(X)$ is evaluated in the Table 4.3. From the column right next to the "Setting" column, we show the unweighted measures. Measures with "$S1$" in the setting are based on validation samples with the same risk factors distribution as the respective training cohorts, while those with "$V1$" in the setting have no verification bias as training and validation cohorts use the same verification mechanism.

In the setting $S1, V1$, the unweighted metrics (C_0 and AUC_0) correspond to the values from internal validation since the training and validation cohorts are generated following the same distributions for risk factors and verification probability. Under this setting, the risk model is well calibrated with all kinds of estimated calibration ratios around 1.03 regardless of being weighted or not, and the unweighted and weighted AUCs are all around 0.75. When the risk factors distributions are the same ($S1$) but the verification probabilities differ ($V2$ and $V3$) between training and validation cohorts, the unweighted metrics and selection weighted metrics are greatly biased with poorly estimated calibration ratios, where the calibration ratios are around 2.9. The AUC is around 0.62 for $S1, V2$ and just 0.47 for $S1, V3$. Once involving the verification weights, the weighted calibration ratios improve substantially with the values around 1.03 no matter with only verification weighting or combined selection and verification weighting. However, when the model for verification weights omits the important variable Z_1 (v_3), the verification weighted or combined weighted calibration ratios are around 1.08, though improved (see $S1, V3$). After weighting for verification bias, the resulting weighted AUCs are all around 0.75 regardless of with only verification weighting or combined weighting.

On the other hand, when the verification mechanisms are the same ($V1$) but the distributions of risk factors differ between training and validation cohorts ($S2$ and $S3$), adjusting only the verification bias or no adjustment at all results in biased estimated verification weighted calibration ratios, i.e. C_0^V be around 1.54 for $S2, V1$ while 1.13 for $S3, V1$. The unweighted and verification weighted AUCs are slightly biased, compared to the reference value 0.75. While adjusting with richest models for weights, w_2v_2, gives the combined weighted calibration ratios $C_0^{WV} = 1.03$ and $C_0^{WV} = 1.02$, respectively for $S2, V1$ and $S3, V1$, while the corresponding AUCs are around 0.75. When the distribution of risk factors for the validation cohort differ in the covariant matrix compared to that in the training ($S3$) and using a model for the selection weights ignoring Z_1, i.e. using w_3, the combined weighted AUCs are biased with the values around 0.80, no matter what verification probability model is used.

When the risk factors distributions and the verification mechanism differ between cohorts, adjusting for either verification bias or selection bias yields severe biased estimated metrics. However, when the combined verification and selection weighting with the models for weights including all important variables (X and Z_1), the resulting weighted metrics are improved. For example, in the setting $S2, V3$, the combined weighted calibration ratios with w_2v_1 or w_2v_2 are around 1.03. But if weighting with w_3v_1 or w_3v_2 instead, i.e. ignoring Z_1 in the model for selection weights, the resulting weighted calibration ratios are poor around 1.20. The combined weighted AUCs are all around 0.75 in settings $S2, V2$ and $S2, V3$. In the setting $S3, V2$ and $S3, V3$, if the model for selection weights ignore Z_1 (w_3), the combined weighted AUCs are biased with values around 0.80, compared to the combined weighted AUCs with selection weights w_1 or w_2 with values ranging from 0.73 to 0.75. The respective combined weighted calibration ratios using w_3 are also biased with values around 1.3 regardless of the models for v. The $S3, V2$ and $S3, V3$ perform poorly with w_3 compared to the corresponding scenarios with $S2$ may be due to the distributions of risk factors used by the validation cohort in $S3$ differ from the training cohort in the covariance matrix, which is difficult to capture by a mis-specified model with missing important predictor for selection bias adjustment.

Based on the simulation results, the proposed weighting method can substantially reduce selection and verification bias in the calculation of the calibration ratio and AUC, even with weighting models that do not include all the factors affecting selection or verification. A rich model that includes main effects, interactions, and quadratic terms of all the relevant factors is especially effective for bias correction in all examined settings.

We also explore the situation if the risk model R is correctly specified including the main effects of both risk factors X and Z_1, i.e. $R(X, Z_1)$, as sensitivity analysis. Table 4.4 summarizes the corresponding results, which are similar as before when the risk model R is not correctly specified including only X.

4.4 Prostate cancer example

We now use a real data sample to illustrate the proposed methods, where we will develop and validate a risk model predicting the 5-year risk of prostate cancer. We use the data from two prostate cancer screening and prevention trials held in North America.

Training and validation data

We use the data from the prostate cancer screening arm of the Prostate, Lung, Colorectal, and Ovarian Trial (PLCO) [Andriole et al., 2009] to develop a 5-year prostate cancer risk prediction model. Men in PLCO had to be 55-74 years old at enrollment, and underwent annual prostate-specific antigen (PSA) testing for six years and annual digital rectal examination (DRE) screening for four years. We use the Selenium and Vitamin E Cancer Prevention Trial (SELECT) to validate the model developed on PLCO [Lippman et al., 2009]. SELECT was a randomized study evaluating the effect of selenium and/or vitamin E supplementation for prostate cancer prevention. Participants in SELECT had to be at least 55 years old if non-African American, and 50 if African American. In contrast to PLCO, men in SELECT were required to have the PSA $\leq$ 4 ng/mL and a normal DRE at enrollment to rule out potential prostate cancer [Cook et al., 2005]. There was no mandatory PSA and DRE annual screening in SELECT, but rather recommended visits at local clinics following community standards every half year. The different PSA screening schedules are also reflected in the left panel of Figure 2.8 plotting the Kaplan-Meier estimate for time to first biopsy, where PLCO biopsied more than SELECT. We exclude patients from both studies with a prior diagnosis of any cancer or missing values of any candidate risk factors.

Table 4.3: Simulation results with risk model $R(X)$. Mean values with standard deviations below over $B = 500$ results calculated with validation cohorts. The risk model R includes the main effects of X. Weights $w_1(v_1)$ are from model including (X, Z_1), $w_2(v_2)$ including (X, Z_1, Z_2), and $w_3(v_3)$ including (X, Z_2). The selection and verification weights, w_i, and v_i, $i = 1, 2, 3$, are based on logistic models that include the main effects, interactions, and quadratic terms. Selection mechanism mechanism in training cohort follows scenario S1. Selection mechanism in validation cohort is from scenarios S1, S2, S3. The verification mechanism in both cohorts is from V1, V2, V3.

Setting	C_0	**Weighted Calibration Measures** C_0^W, C_0^V, C_0^{WV}														
		w_1	w_2	w_3	v_1	v_2	v_3	w_1v_1	w_1v_2	w_1v_3	w_2v_1	w_2v_2	w_2v_3	w_3v_1	w_3v_2	w_3v_3
S1, V1	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03	1.03
	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
S1, V2	2.93	2.92	2.92	2.92	1.02	1.02	1.04	1.02	1.02	1.04	1.02	1.02	1.04	1.02	1.02	1.04
	0.14	0.14	0.14	0.14	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
S1, V3	2.90	2.90	2.90	2.90	1.03	1.03	1.08	1.03	1.03	1.07	1.03	1.03	1.07	1.03	1.03	1.08
	0.15	0.15	0.15	0.15	0.10	0.10	0.11	0.10	0.10	0.11	0.10	0.10	0.11	0.10	0.10	0.11
S2, V1	1.53	1.07	1.02	1.20	1.54	1.53	1.54	1.08	1.07	1.08	1.03	1.03	1.04	1.20	1.20	1.21
	0.10	0.08	0.08	0.09	0.10	0.10	0.10	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.09	0.09
S2, V2	4.44	3.06	2.92	3.45	1.52	1.51	1.56	1.07	1.06	1.08	1.02	1.02	1.04	1.20	1.20	1.23
	0.17	0.13	0.14	0.16	0.06	0.06	0.07	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.06
S2, V3	4.50	3.03	2.90	3.46	1.53	1.52	1.61	1.08	1.07	1.10	1.03	1.03	1.06	1.20	1.20	1.27
	0.18	0.14	0.14	0.16	0.12	0.12	0.12	0.10	0.10	0.10	0.10	0.10	0.10	0.11	0.11	0.11
S3, V1	1.12	0.90	1.01	1.25	1.13	1.13	1.14	0.91	0.91	0.91	1.02	1.02	1.01	1.26	1.26	1.26
	0.09	0.12	0.19	0.16	0.09	0.09	0.09	0.11	0.12	0.12	0.19	0.19	0.18	0.16	0.15	0.16
S3, V2	2.87	2.69	2.92	3.06	1.12	1.12	1.16	0.90	0.90	0.92	1.02	1.01	1.03	1.25	1.25	1.28
	0.14	0.24	0.38	0.20	0.06	0.06	0.06	0.08	0.08	0.08	0.15	0.15	0.15	0.13	0.13	0.13
S3, V3	2.50	2.81	2.94	2.53	1.14	1.14	1.20	0.92	0.92	0.95	1.03	1.02	1.05	1.27	1.27	1.34
	0.13	0.38	0.54	0.17	0.11	0.11	0.12	0.17	0.17	0.15	0.23	0.23	0.23	0.27	0.26	0.28
Scenario	AUC_0	**Weighted AUC Estimates** $AUC_0^W, AUC_0^V, AUC_0^{WV}$														
		w_1	w_2	w_3	v_1	v_2	v_3	w_1v_1	w_1v_2	w_1v_3	w_2v_1	w_2v_2	w_2v_3	w_3v_1	w_3v_2	w_3v_3
S1, V1	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S1, V2	0.62	0.62	0.62	0.62	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S1, V3	0.47	0.47	0.47	0.47	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S2, V1	0.76	0.75	0.75	0.76	0.76	0.76	0.76	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S2, V2	0.62	0.62	0.62	0.62	0.76	0.76	0.76	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S2, V3	0.48	0.47	0.47	0.48	0.76	0.76	0.76	0.75	0.75	0.76	0.75	0.75	0.75	0.75	0.75	0.75
	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S3, V1	0.79	0.74	0.75	0.81	0.79	0.79	0.80	0.73	0.73	0.74	0.75	0.75	0.75	0.81	0.81	0.81
	0.02	0.03	0.04	0.02	0.02	0.02	0.01	0.03	0.04	0.03	0.04	0.04	0.04	0.02	0.02	0.02
S3, V2	0.67	0.60	0.62	0.70	0.79	0.79	0.79	0.73	0.73	0.74	0.75	0.75	0.75	0.81	0.81	0.81
	0.01	0.03	0.03	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02
S3, V3	0.54	0.45	0.48	0.56	0.79	0.79	0.79	0.73	0.73	0.73	0.74	0.74	0.74	0.80	0.80	0.81
	0.01	0.04	0.04	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.04	0.03	0.03	0.03

Table 4.4: Simulation results with correctly specified risk model, $R(X, Z_1)$. Mean values with standard deviations below over $B = 500$ results calculated with validation cohorts. Risk model R includes the main effects of X and Z_1. Weights $w_1(v_1)$ are from model including (X, Z_1), $w_2(v_2)$ including (X, Z_1, Z_2), and $w_3(v_3)$ including (X, Z_2). The selection and verification weights are based on logistic models that included the main effects, interactions, and quadratic terms. Selection mechanism mechanism in training cohort follows scenario S1. Selection mechanism in validation cohort is from scenarios S1, S2, S3. The verification mechanism in both cohorts is from V1, V2, V3.

Scenario	C_0	**Weighted Calibration Measures** C_0^W, C_0^V, C_0^{WV}														
		w_1	w_2	w_3	v_1	v_2	v_3	w_1v_1	w_1v_2	w_1v_3	w_2v_1	w_2v_2	w_2v_3	w_3v_1	w_3v_2	w_3v_3
S1, V1	1.04	1.03	1.03	1.03	1.04	1.04	1.03	1.04	1.04	1.03	1.04	1.04	1.03	1.04	1.04	1.03
	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
S1, V2	2.93	2.92	2.93	2.93	1.02	1.02	1.04	1.02	1.02	1.04	1.02	1.02	1.04	1.02	1.02	1.04
	0.14	0.14	0.14	0.14	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
S1, V3	2.91	2.91	2.91	2.91	1.03	1.03	1.08	1.03	1.03	1.08	1.03	1.03	1.08	1.03	1.03	1.08
	0.15	0.15	0.15	0.15	0.10	0.10	0.11	0.10	0.10	0.11	0.10	0.10	0.11	0.10	0.10	0.11
S2, V1	1.07	1.07	1.03	1.00	1.08	1.07	1.08	1.08	1.08	1.09	1.03	1.03	1.04	1.01	1.01	1.02
	0.07	0.08	0.08	0.08	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
S2, V2	3.10	3.05	2.92	2.88	1.06	1.06	1.09	1.07	1.06	1.08	1.02	1.02	1.04	1.00	1.00	1.03
	0.12	0.13	0.14	0.13	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.05	0.05	0.05
S2, V3	3.15	3.04	2.91	2.89	1.07	1.07	1.13	1.08	1.08	1.10	1.03	1.03	1.07	1.01	1.01	1.07
	0.13	0.14	0.14	0.13	0.08	0.08	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.09
S3, V1	0.86	0.90	1.01	0.94	0.87	0.87	0.88	0.91	0.91	0.91	1.02	1.02	1.02	0.95	0.95	0.96
	0.07	0.11	0.19	0.12	0.07	0.07	0.07	0.11	0.11	0.11	0.18	0.18	0.18	0.12	0.12	0.12
S3, V2	2.21	2.68	2.91	2.31	0.86	0.86	0.89	0.90	0.90	0.92	1.01	1.01	1.03	0.94	0.94	0.97
	0.11	0.23	0.37	0.16	0.05	0.05	0.05	0.08	0.08	0.08	0.15	0.15	0.15	0.10	0.09	0.10
S3, V3	1.93	2.80	2.94	1.92	0.88	0.88	0.93	0.91	0.91	0.95	1.02	1.02	1.05	0.97	0.96	1.01
	0.10	0.36	0.50	0.14	0.09	0.09	0.09	0.17	0.17	0.15	0.23	0.23	0.22	0.20	0.20	0.21
Scenario	AUC_0	**Weighted AUC Estimates** $AUC_0^W, AUC_0^V, AUC_0^{WV}$														
		w_1	w_2	w_3	v_1	v_2	v_3	w_1v_1	w_1v_2	w_1v_3	w_2v_1	w_2v_2	w_2v_3	w_3v_1	w_3v_2	w_3v_3
S1, V1	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S1, V2	0.72	0.72	0.72	0.72	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S1, V3	0.63	0.63	0.63	0.63	0.80	0.80	0.81	0.80	0.80	0.81	0.80	0.80	0.81	0.80	0.80	0.81
	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S2, V1	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
S2, V2	0.73	0.72	0.72	0.72	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
S2, V3	0.63	0.63	0.63	0.63	0.80	0.80	0.81	0.80	0.80	0.81	0.80	0.80	0.81	0.80	0.80	0.81
	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.01
S3, V1	0.81	0.78	0.80	0.82	0.81	0.81	0.81	0.78	0.78	0.78	0.80	0.80	0.80	0.82	0.82	0.83
	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.02	0.02	0.02
S3, V2	0.72	0.69	0.72	0.73	0.81	0.81	0.81	0.78	0.78	0.78	0.80	0.80	0.80	0.82	0.82	0.83
	0.01	0.02	0.03	0.02	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02
S3, V3	0.60	0.60	0.63	0.62	0.81	0.81	0.82	0.78	0.78	0.79	0.80	0.80	0.81	0.82	0.82	0.83
	0.02	0.03	0.04	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03

Because not all participants had a PSA value measured at study entry, we define the baseline as the age of the first PSA measurement in this application. We exclude SELECT men younger than 55 years and older than 74 years at baseline since older men are not typically screened for prostate cancer as indicated by exclusion from PLCO. We further exclude PLCO participants with PSA > 10 ng/mL at baseline as these are typically referred to the urologists and exempt from screening tools [Ankerst et al., 2014]. These exclusion criteria result in N_1 = 29699 men in the PLCO training set and N_0 = 26422 in the SELECT validation set described in Table 2.2.

Men in PLCO tended to receive biopsies earlier than in SELECT as they were mandatory to screen for PSA and DRE, which causes a shorter time to prostate cancer diagnosis as shown in the right panel of Figure 2.8. The steep increase in the curve for PLCO in the first six months after the first PSA measurement reflects the detection of prevalent prostate cancers. In contrast, SELECT excluded cancer cases at study entry by requiring a normal PSA and DRE at enrollment.

Risk model estimation

We build a risk model for prostate cancer based on the PLCO training data using a Cox proportional hazards model with age as the underlying time metric because it is more reasonable to expect the baseline hazard for prostate cancer is a function depending on age rather than on the time in trials. The hazard function is defined as $\lambda(t|\mathbf{X}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{X})$, where $\lambda_0(t)$ is the unspecified baseline hazard function, $\boldsymbol{\beta}$ is the vector of log relative risk parameters and $\mathbf{X}$ denotes the baseline risk factors [Cox, 1972]. Let Y^* denote the age at prostate cancer diagnosis, L_j denote the age at study entry, i.e. age at first PSA test, and C_j the age at censoring for individual $j = 1, \ldots, N_1$. The observed age for the jth individual is $Y_j = \min\left(Y_j^*, C_j\right)$ and $\delta_j = I(L_j < Y_j^* \leq C_j)$ is his censoring indicator. We assume that L_j and C_j are independent of Y_j^* given $\mathbf{X}_j$ and participants are censored at 5 years after baseline when estimating the model.

We obtain the log-hazard ratio estimates $\hat{\boldsymbol{\beta}}$ from the standard Cox partial likelihood described in Section 1.1 and use the Bayesian information criterion (BIC) to select the predictors, allowing up to two-way interactions of the first PSA value in log-base-2 scale ($log_2 PSA$), family history, African ancestry, and prior negative biopsy. Table 4.5 summarizes the estimated hazard ratios of the final Cox model. A higher PSA, having a first-degree family history of prostate cancer, and a prior negative biopsy all significantly increase the risk of being diagnosed with prostate cancer within five years after baseline under annual screening (all $p < 0.001$ except for prior negative biopsy with $p = 0.008$). Among the three risk factors, the PSA increases the cancer risk the most as expected. The interaction term indicates that for men with prior negative biopsy, the prostate cancer risk diminishes with increasing PSA, possibly due to a benign condition, such as having benign hyperplasia may contribute to the increasing PSA rather than prostate cancer.

Given $\hat{\beta}$ and the Breslow estimate of the cumulative baseline hazard $\hat{\Lambda}_0(a)$ at age a plotted in Figure 4.1 with ranges from 0 to around 0.07 when age ranging from 55 to 79 [Breslow, 1972], the predicted τ–year risk of prostate cancer for a man with risk factors $\mathbf{X}$ and baseline age a is calculated by

$$R\left(\tau, \mathbf{X}, a\right) = P\left(Y^{*} \leq \tau + a \,|\mathbf{X}, Y^{*} > a\right) = 1 - \exp\left[-\{\widehat{\Lambda}_0(\tau + a) - \widehat{\Lambda}_0(a)\}\exp\left(\widehat{\boldsymbol{\beta}}'\mathbf{X}\right)\right].$$

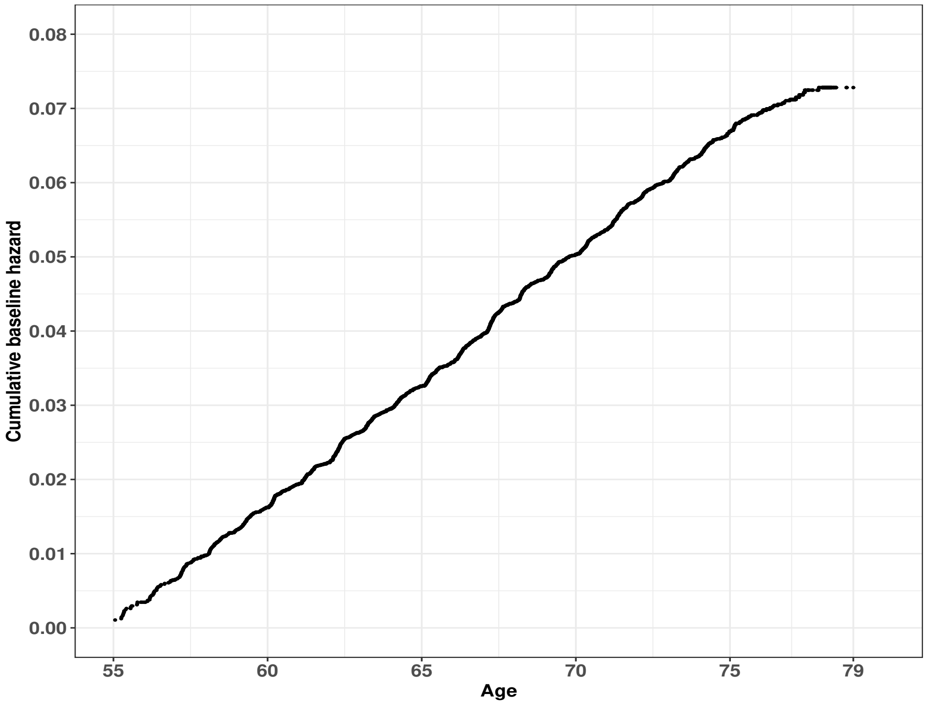

Figure 4.1: Cumulative baseline hazard function from the Cox model for 5-year prostate cancer risk estimated based on 29699 PLCO participants.

Table 4.5: Hazard ratios and 95% confidence intervals (CIs) estimated from the Cox model for 5-year prostate cancer risk fit to 29699 PLCO participants using age as the time metric. All risk factors have p-values $<$ 0.001 except for $Prior\ negative\ biopsy$, which has a p-value of 0.008.

Risk factor	Hazard ratio (95% CI)
$log_2 PSA$	4.32 (4.09, 4.58)
$Family\ history$	1.42 (1.22, 1.65)
$Prior\ negative\ biopsy$	1.83 (1.17, 2.87)
$log_2 PSA * Prior\ negative\ biopsy$	0.61 (0.50, 0.74)

To illustrate the range of estimated probabilities from the model, we show the estimated 5-year prostate cancer risk for several profiles in Figure 4.2 based on the following combinations of risk factors: age at baseline be 55, 66, or 74, family history be yes or no, no prior negative biopsy, and different baseline PSA values. In general, the cancer risk increase when the PSA level increase in each combination. Within groups of the same age, men with no family history have a lower prostate cancer risk than those with family history. Among men without a family history of cancer, the prostate cancer risk for men at 74 years old is the lowest among the three different age levels, while at 65 is the highest. This may be due to fewer elderly participants joining the trials. According to Table 2.2, men with age between 70 to 74 years only consist a small proportion in respective cohorts (11.0% in PLCO; 10.3% in SELECT).

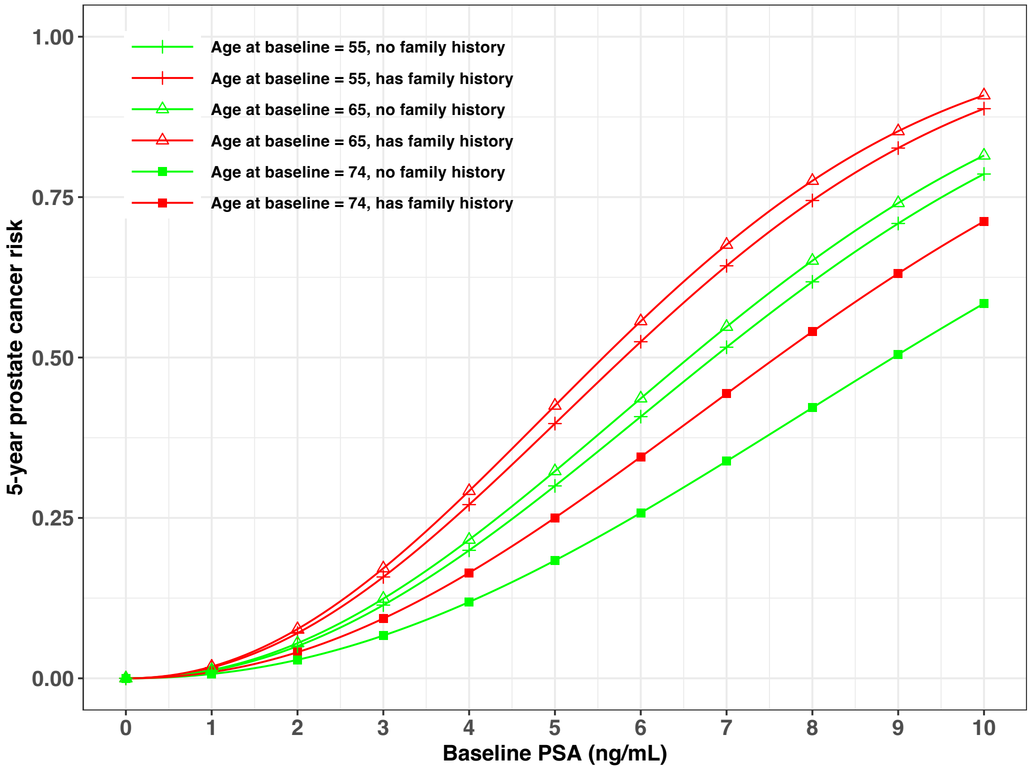

Figure 4.2: Estimated 5-year prostate cancer risk from the risk model built on 29699 PLCO participants for select risk profiles with various ages at baseline, with/without family history of prostate cancer, and no prior negative biopsy.

Accounting for censoring

Censoring occurs when the follow-up ends before the projection time τ, due to events other than the outcome of interest. One approach to accommodate censoring is to compute R_i for those who are censored before τ only up to their censoring time and set $D_i = 0$ [Pfeiffer and Gail, 2017]. This approach yields unbiased estimates of calibration, but is based on variate projection intervals rather than a fixed time τ. Another method suggested by Li et al. [2018] is to impute the outcome for those who censored before τ. In their method, let Y_i denote the observed survival time and δ_i the event indicator (1 if event, 0 if censored) for subject i in the validation cohort, the censoring weighted outcome for individual i is

$$\begin{aligned} D_i^C(\tau) &= P(T_i \leq \tau | R_i, \delta_i, Y_i) = E[D_i(\tau)|R_i, \delta_i, Y_i] \\ &= \left[1 - (1-\delta_i)\frac{S_T(\tau|R_i)}{S_T(Y_i|R_i)}\right] I\{Y_i \leq \tau\}. \end{aligned} \tag{4.78}$$

The conditional survival function $S(t|R_i)$ is estimated using a kernel weighted Kaplan-Meier estimate

$$\hat{S}(t|R_i) = \hat{P}(T_i \geq t|R_i) = \prod_{s\in\Omega, s\leq t} \left\{ 1 - \frac{\sum\limits_j K_h(R_j, R_i) I(Y_j = s)\delta_j}{\sum\limits_j K_h(R_j, R_i) I(Y_j \geq s)} \right\} \tag{4.79}$$

where $K_h(R_j, R_i) = I\{|(R_j - R_i)/h| \leq 1\}/2h$ is a kernel weight with band width h and Ω denotes the set of distinct event times. We estimate $K_h(R_j, R_i)$ with the function *calc.kw* in the *R* package *tdROC* with bandwidth $h = 0.003$ and $\widehat{S}(t|R_i)$ using *survfit* from the *survival* package with K_h as the weight. When apply the method by Li et al. [2018], the censoring weighted outcome D^C is 1 for men with cancer within 5 years, 0 for men who survive longer than 5 years, while some probability values for the rest. We evaluate the performance measures using the censoring weighted outcome, D^C, in the sensitivity analysis.

Bootstrapping procedure

We use bootstrapping to estimate the 95% percentiles confidence intervals in this application, which procedure is described in Algorithm 1. In each of the B bootstrap repetitions, we generate new PLCO and SELECT by sampling with replacements from the PLCO and SELECT cohorts, respectively. The models for the probability of being in PLCO, the probability of being verified in PLCO, and the probability of being verified in SELECT are then refitted using the new bootstrap samples. If the censoring outcome as defined in (4.78) is used to calculate the performance measure, we also re-estimate the D^C for the new SELECT sample. At last, we calculate the unweighted and weighted performance measure using the new SELECT sample, where the weights come from the refitted models. We then construct the 95% confidence interval for each performance measure using the corresponding 2.5% and 97.5% percentiles of the bootstrapping output.

Internal validation of the prediction model

To assess model performance in PLCO, we apply five-fold cross-validation by dividing the PLCO cohort into five non-overlapping subsets of equal size. To avoid numerical problems, we ensure that roughly equal numbers of prostate cancer cases are included in all subsets. We successively use four subsets to estimate the risk model R in (4.78) and the remaining

Algorithm 1: Bootstrapping procedure.

```
input  : PLCO, SELECT
output: Estimated unweighted and weighted performance measures
for $i = 1, \ldots, B$ do
    (1) draw PLCO$_i$ (SELECT$_i$) from PLCO (SELECT) with replacement such that
      PLCO$_i$ (SELECT$_i$) has the same size as PLCO (SELECT).
    (2) refit the model for the probability of being in PLCO with combined data of
      PLCO$_i$ and SELECT$_i$.
    (3) refit the model of being verified in PLCO with PLCO$_i$ (with step-wise selection
      if applicable).
    (4) refit the model of being verified in SELECT with SELECT$_i$ (with step-wise
      selection if applicable).
    (5) applied the refitted models to SELECT$_i$ calculating the selection, verification,
      and combined weights.
    if using censoring outcome then
        (6) re-calculate the censoring outcome ($D^C$) for SELECT$_i$.
        (7) estimate the unweighted and weighted performances measures with
          SELECT$_i$ using the new weights and $D^C$.
    end
    if using observed outcome $D$ then
        (6) estimate the unweighted and weighted performances measures with
          SELECT$_i$ using the new weights.
    end
end
```

subset to validate the model. The average of these five performance measures is the cross-validated estimates (see Algorithm 2 for details).

The C_1 and AUC_1 in 29699 PLCO participants estimated using cross-validated predicted risks with D are $C_1 = 0.993$ (95% CI: $0.951, 1.035$) and $AUC_1 = 0.883$ (95% CI: $0.875, 0.892$) as given in Table 4.6, indicating the model is well calibrated and has a good discriminating ability. The results using D and D^C are similar, with the C_1 using D^C being around 1.003. Confidence intervals are based on the 2.5% and 97.5% percentiles from bootstrapping with 600 repetitions of the cross-validated predictions, which procedure is described in Algorithm 3. Here, we simply calculate the performance measures with the bootstrap sample without re-estimating the cancer risk in each repetition due to the concern that refitting the model and re-estimating the cancer risk may result in too much variation.

Table 4.6: Estimated unweighted calibration ratio (C_1) and area under the receiver operating characteristic curve (AUC_1) in 29699 PLCO participants, using D or censoring outcome D^C. 95% confidence intervals (CIs) are percentiles of the bootstrap empirical distribution function with 600 bootstrap repetitions.

	Estimate with D (95% CI)	Estimate with D^C (95% CI)
C_1	0.993 (0.951, 1.035)	1.003 (0.961,1.045)
AUC_1	0.883 (0.875, 0.892)	0.883 (0.875, 0.892)

Algorithm 2: Five-fold cross-validation procedure

```
input  : PLCO
output: Estimated unweighted performance measures

(1) split PLCO into five disjoint subsets.
for $i = 1, \dots, 5$ do
    (a) train the model for cancer risk in Table 4.5 with subsets $j, j \neq i$.
    (b) apply model from (a) to subset $i$ estimating the cancer risk.
    (c) calculate the unweighted performance measures with subset $i$.
end
(2) output the average of the five estimated unweighted performance measures.
```

Algorithm 3: Bootstrapping procedure for internal validation

```
input  : PLCO with estimated cancer risk from five-fold cross-validation
output: Estimated unweighted performance measures

for $i = 1, \dots, B$ do
    (1) draw PLCO$_i$ from PLCO with replacement such that PLCO$_i$ has the same
      size as PLCO.
    (2) calculate the unweighted performance measures using PLCO$_i$.
end
```

External model validation

Before validating the developed prostate cancer risk prediction model with SELECT, we first build the models for the selection and verification weights, where different logistic regressions are estimated. We then apply the model for cancer risk and the models for weights to the SELECT cohort and calculate the unweighted and weighted performance measures, where we also explore the situation when using censoring outcome instead in the calculation of performance measures.

Estimation of the selection and verification weights

We fit a logistic regression model to the combined PLCO and SELECT data estimating $P(T = 1|\mathbf{X}^*)$ to obtain $\hat{w}(\mathbf{X}^*)$ in (4.11), where $T = 1$ for PLCO and $T = 0$ for SELECT. Step-wise model selection with the BIC selection criterion is used, allowing up to two-way interactions of age at first PSA test, log-based-2 PSA ($log_2 PSA$), family history, African ancestry, and prior negative biopsy. Table 4.7 describes the estimated odds ratios for the final logistic model for the probability of being in PLCO. Men with lower PSA values, having a family history, having African ancestry, or having a prior negative biopsy are more likely to be in SELECT than in PLCO. In sensitivity analysis, we estimate the selection weights using the same logistic model but without model selection (see model described in Table 4.10).

Verification in this application refers to receiving a prostate biopsy within 5 years after baseline. To obtain $\hat{v}(\mathbf{X}^*)$, we fit two logistic regression models as in (4.53) to estimate the probabilities of being verified in PLCO or SELECT, respectively. The values of PSA and DRE used in these logistic models are those closest to the first biopsy for verified participants or closest to the end of the five-year follow-up after baseline for unverified participants. Table 4.8 shows the odds ratios of the resulting logistic regression models for PLCO and SELECT. The estimated odds ratio for PSA $>$ ng/mL and abnormal DRE are much higher than the other risk factors, which is due to that both trials were recommended to biopsy men with PSA $>$ ng/mL and abnormal DRE. These two odds ratios in PLCO are higher than the respective ones in SELECT. Having a family history of cancer increases the risk of being biopsied in both trials, while having a prior negative biopsy history reduces the risk of being biopsied in PLCO but increases in SELECT, highlighting the differences in disease verification mechanism between the two cohorts.

Figure 4.3 presents the estimated SELECT 5-year verification risk $\hat{P}_0(V = 1|\mathbf{X}^*)$ on the x-axis versus the PLCO 5-year verification risk $\hat{P}_1(V = 1|\mathbf{X}^*)$ on the y-axis for the 26422 SELECT participants. From Figure 2.8, it appears that verification is more frequent in PLCO as the cumulative incidence curve for biopsy for PLCO is higher than the one for SELECT indicating more biopsy events in PLCO. However, after conditioning on risk factors, the 5-year SELECT verification probability ($P_0(V = 1|\mathbf{X}^*)$) is on average higher than the corresponding 5-year PLCO verification probability ($P_1(V = 1|\mathbf{X}^*)$) for most men in SELECT, resulting in most values of v being much smaller than one. In sensitivity analysis, we add the number of PSA tests within 5 years after baseline to the logistic models (see characteristic description in Table 4.11) and apply different model selection processes to model $P_T(V = 1|\mathbf{X}^*), T = 0, 1$ (see Table 4.13). The percentages of participants with PSA $>$ 4 ng/mL and abnormal DRE are higher than those in SELECT. The distributions of the number of PSA tests within 5 years after baseline are comparable between the two cohorts. PLCO has much more men who received four tests compared to SELECT (21.1% versus 13.7%). SELECT has 15 participants taken more than 7 PSA tests, while none in PLCO. Table 4.12 summarise the resulting models for verification probabilities under various model selection processes.

Table 4.7: Odds ratios and 95% confidence intervals from the step-wise logistic selection weight model with outcome (1:in PLCO versus 0: in SELECT) applied to the 56121 participants of both studies (29699 from PLCO and 26422 from SELECT). All risk factors have p-values $<$ 0.001 except for log_2PSA, which has a p-value of 0.008. PSA = prostate-specific antigen.

Risk factor	Odds ratio	95% Confidence interval
$Intercept$	2.14	(1.72, 2.65)
log_2PSA	0.78	(0.64, 0.94)
Age	0.993	(0.990, 0.997)
$Family\ history$	0.39	(0.37, 0.41)
$African\ ancestry$	0.33	(0.30, 0.35)
$Prior\ negative\ biopsy$	0.43	(0.40, 0.47)
$log_2PSA * Age$	1.006	(1.003, 1.009)
$log_2PSA * African\ ancestry$	1.15	(1.08, 1.23)
$log_2PSA * Prior\ negative\ biopsy$	1.18	(1.11, 1.25)

Table 4.8: Odds ratios and 95% confidence intervals (CIs) from logistic models fit to having a biopsy within 5 years fit to 29699 PLCO and 26422 SELECT participants respectively. PSA = prostate-specific antigen; DRE = digital rectal exam.

	PLCO (N_1 = 29699)	SELECT (N_0 = 26422)
Risk factor	Odds ratios (95% CI)	Odds ratios (95% CI)
$Intercept$	0.005 (0.002, 0.014)	0.015 (0.013, 0.017)
$log_2 PSA$	4.73 (3.73, 6.03)	4.46 (4.06, 4.92)
$I(PSA > 4)$	350.38 (93.18,1320.18)	28.20 (15.32, 50.92)
$DRE(abnormal)$	511.01 (364.03, 735.38)	86.35 (69.22, 107.94)
Age	0.98 (0.97, 1.00)	-
$Family\ history$	1.22 (1.03, 1.46)	1.37 (1.21, 1.54)
$African\ ancestry$	0.010 (0.001, 0.219)	-
$Prior\ negative\ biopsy$	0.81 (0.68, 0.97)	1.77 (1.33, 2.34)
$log_2 PSA * I(PSA > 4)$	0.78 (0.60, 1.01)	0.38 (0.29, 0.49)
$log_2 PSA * DRE(abnormal)$	0.27 (0.21, 0.35)	0.31 (0.26, 0.37)
$log_2 PSA * Prior\ negative\ biopsy$	-	0.76 (0.65, 0.89)
$I(PSA > 4) * DRE(abnormal)$	0.11 (0.07, 0.17)	0.62 (0.35, 1.12)
$I(PSA > 4) * Age$	0.97 (0.95, 0.99)	-
$Age * African\ ancestry$	1.08 (1.02, 1.13)	-

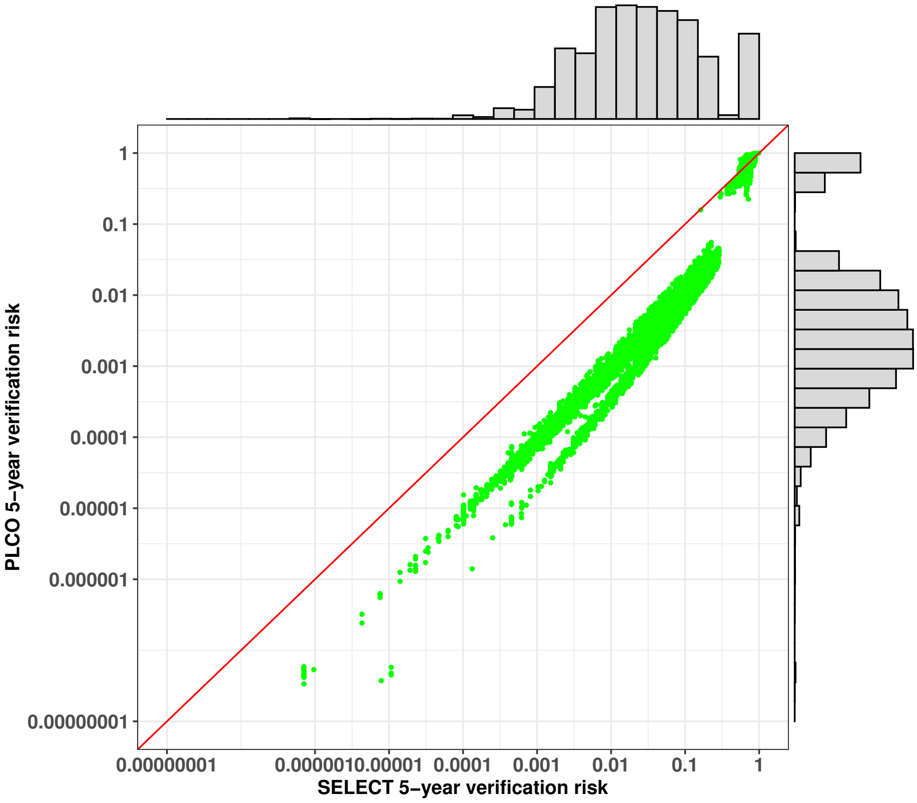

Figure 4.3: Estimated SELECT 5-year verification probability $\hat{P}_0(V = 1|\mathbf{X}^*)$ on the x-axis versus PLCO 5-year verification probability $\hat{P}_1(V = 1|\mathbf{X}^*)$ on the y-axis estimated from logistic models given in Table 4.8 on 26422 SELECT participants. The marginal histogram of the SELECT verification probability is given on top of the scatter plot, the PLCO verification risk histogram is given to the right. The diagonal line indicates equal probabilities.

Performance measures

Table 4.9 gives the estimated unweighted and weighted calibration ratios and AUCs in 26422 SELECT participants. Confidence intervals are based on the 2.5% and 97.5% percentiles from the bootstrapping with 600 repetitions following the procedure shown in Algorithm 1. Such bootstrapping procedure re-samples individuals from PLCO and SELECT and also re-estimates the weights in each bootstrap repetition, thus accounting for all the uncertainties in the performance estimates.

The overall unweighted calibration ratio $C_0 = 1.191$ suggests that the model under-predicts the number of events in SELECT since the ratio is over one. The selection weighted calibration ratio is $C_0^W = 1.155$. Though still pronounced, the selection bias adjustment reduces the under-prediction by around 0.035. Figure 4.4 provides further insight into unweighted model performance and after accommodating for selection bias in the subgroups of risk factors among the 26422 SELECT participants. The estimated selection weights for participants with family history, African ancestry, and prior negative biopsy differ largely from the corresponding rest groups, while similar when comparing the subgroup with first PSA $>$ 2 ng/mL versus without or the subgroup with age at baseline $>$ 65 versus without. The estimated unweighted calibration ratio is worst in the subgroup with African ancestry. But this subgroup also receives lower selection weights compared to the subgroup with no African ancestry, giving an insight into the bias reduction after weighting though the resulting calibration ratio is based on overall calculation. The reduction in the selection weighted calibration ratio may attribute to giving lower weights to those with poorer calibration performance.

Adjusting verification bias alone leads to $C_0^V = 0.893$, corresponding to an over-prediction of the model after accommodating differences in prostate cancer verification mechanism. Adjusting the selection bias in additional does not improve the calibration further with $C_0^{WV} = 0.884$. Figure 4.5 shows the distributions of verification weights and the corresponding performance measure C_0^V in various risk factors subgroups. Only a small proportion of men in SELECT have verification weights $v > 1$, which has been indicated by Figure 4.3. After verification weighting, the calibration ratio in the subgroup with African ancestry reduces from the unweighted value 1.736 to 1.336, while from the 1.339 to 0.939 for the subgroup with family history. Figure 4.6 shows the distributions of the combined selection and verification weights and C_0^{WV} by risk factors subgroups. Including selection weights additionally does not noticeably change the weighted calibration measures in any risk factor subgroup, compared to those with only verification weighting. The vastly different values of the verification weighted calibration ratios, overall or in subgroups of risk factors, emphasizing the need to account for the differences in the verification mechanism between the validation and training cohorts.

The selection weighted AUC ($AUC_0^W = 0.824$) is just slightly lower than the unweighted one ($AUC_0 = 0.828$), but the verification weighted AUC increases to $AUC_0^V = 0.853$. Combined selection and verification weighting do not further improve the AUC ($AUC^{WV} = 0.851$). However, including the adjustment for verification bias improves the resulting estimated AUC compared to the unweighted one, indicating the need to accommodate the difference in the verification process between the training and validation populations. Figure 4.8 presents the estimated unweighted and weighted TPRs and FPRs versus the risk threshold ranging from 0 to 0.3 among 26422 SELECT participants. The FPR values are not changed by weighting since the curves are overlapping, but the verification weighted and combined weighted TPR curves are higher compared to the unweighted or selection weighted ones, which explains the noticeable improvement in the verification weighted and combined weighted AUCs. As shown in Figure 4.8, the corresponding verification weighted or combined weighted ROCs are higher than the other two curves as well.

Table 4.9: Estimated unweighted and weighted calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants. 95% confidence intervals are percentiles of the bootstrap empirical distribution function with 600 bootstrap repetitions. The outcome without censoring weighting (D) is used in the calculation of the measures. The selection weights w are from the model in Table 4.7 and the verification weights v are from the models in Table 4.8.

	Estimate	95% Confidence interval
C_0	1.191	(1.126, 1.258)
C_0^W	1.155	(1.086, 1.221)
C_0^V	0.893	(0.839, 0.952)
C_0^{WV}	0.884	(0.824, 0.941)
AUC_0	0.828	(0.817, 0.840)
AUC_0^W	0.824	(0.812, 0.835)
AUC_0^V	0.853	(0.842, 0.865)
AUC_0^{WV}	0.851	(0.839, 0.862)

Sensitivity analyses

When the selection weights w and verification weights v are based on different models, the resulting estimated performance measures giving in Table 4.13 are similar as before. The selection weighted calibration ratio with w from the model without model selection (MS) is 1.157 versus the 1.155 before. Using models from Table 4.8 for v and w without MS gives the combined weighted calibration ratio of 0.886, versus 0.884 before. When using the selection weights given by model in Table 4.7 but using different models described in Table 4.12 for v, the estimated verification weighted calibration ratios are 0.878, 0.858, and 0.833 for models VM2, VM3, and VM4, respectively. The estimated C_0^V when using the model VM4 for v, i.e. the optimal model after step-wise model selection including additionally the number of PSA tests as a candidate risk factor, is the worst, which means the VM4 may not be able to address the verification bias properly. Similarly, the estimated combined weighted calibration ratio when using VM4 for v is just 0.828 or 0.830 when using w with or without model selection, respectively. The AUC_0^W is similar as before when w from the model without MS. The verification or combined weighted AUCs are comparable without much variation no matter which models are used for w and v, ranging from 0.850 to 0.853.

If using the probability of censoring weighted outcome, D^C instead, the validation results are similar to those based on the censored projection times as shown in Table 4.14. Figures 4.9, 4.10, and 4.11 present the distributions of the estimated weights for various risk factor subgroups with the estimated unweighted or weighted calibration measures using the censoring weighted outcomes D^C given alongside as before. The densities of weights by subgroups also look similar to those using D and therefore, we leave out further interpretation here.

4.5 Discussion

Validating the developed clinical risk models with external data is essential before extensive applications. If the validation population and the training one resemble, the performance measures calculated with the validation data should have a similar value as those from internal validation, i.e. validation with samples from the training population. In this case, one evaluates the “reproducibility” of the developed risk models. On the other hand, when the distributions of risk factors or the disease verification mechanism, or both are different between the training data and the external validation population, we assess rather the “transportability” of the models in the external validation. Such differences in either the risk factors distributions or the disease diagnostic test scheme may distort the estimated performance measure and fallaciously suggest that the developed risk model should not be used though the model is valid for the training population.

We developed a novel weighting method to address the differences in the risk factor distributions or in the disease verification mechanism between training and validation populations in the external validation of clinical risk models. The selection or verification weights are proposed to adjust for selection bias arising from different distributions in risk factors or verification bias arising from different verification mechanisms between training and validation cohorts, respectively. These weights capture the differences and can improve the assessment of model performance after incorporating them into the calculation of the performance measures.

When only the difference in the risk factor distributions is considered, we apply the selection weight to resemble the external sample to the training population. We formalized the concepts of reproducibility and transportability. If the unweighted performance measures have similar values as those from internal validation, the model is reproducible and can be applied to the population similar to the training one. If only the selection weighted, but not the unweighted, performance measures resemble the values of unweighted measures from internal validation, we conclude that the model is transportable to other populations that may not have similar risk factor distributions as the training population. During the discussion of reproducibility and transportability, we always assume that the true disease probabilities are the same between training and validation populations based on predictors in the risk model ($\mathbf{X}$) and other variables omitted in the risk modeling ($\mathbf{Z}$), i.e. $\pi_1(\mathbf{X}^*) = \pi_0(\mathbf{X}^*)$. If the conditional distribution of omitted covariates given predictors is the same between training and validation sets ($F_1(\mathbf{Z}|\mathbf{X}) = F_0(\mathbf{Z}|\mathbf{X})$) in addition, the risk model $R(\mathbf{X})$ is well calibrated with selection weighting in the validation data, i.e. model is transportable. Instead, if the marginal distributions of X are the same in the two distributions ($F_1(\mathbf{X}) = F_0(\mathbf{X})$), $R(\mathbf{X})$ is well calibrated without selection weighting and the model is reproducible in this case. Other approaches to address the heterogeneity in the distribution of risk factors in external validation through weighting include the work by Powers et al. [2019] that used weighting to estimate the performance of a model with external samples to resemble its performance in the target population, whose work had a different aim than ours and the difference in verification scheme was omitted.

Towards the adjustment of bias from different verification mechanisms between cohorts, we proposed the verification weights, where the verified and unverified participants receive different forms of verification weight. For adjusting the verification bias, we weighted the observations but not the predictions. When taking the verification process into account, we rather model the probability of being verified and being tested positive for the disease in the verification given risk factors ($P(DV = 1|\mathbf{X}^*)$), than the disease probability $P(D = 1|\mathbf{X}^*)$. We discussed the situations when the external performance measures have similar values as internal validation after verification weighting as well. When the distributions of predictors and other covariates are the same and the probabilities of disease conditional on $\mathbf{X}^*$ and be verified are the same between the training and validation cohorts, the verification weighted performance measures are similar to those from internal validation. We also gave the expressions for combined weighted performance measures accounting for both selection and validation bias.

The accommodation of the selection bias and verification bias relies on modeling the verification probabilities in training and the validation populations, respectively. Correctly modeling these probabilities is essential for bias correction, which is a challenge as shown in the prostate cancer example that we have intrinsic different screening plans between the two sets and only limited risk factors are available from both sets. The modeling variable selection and the existence of potential confounders could add extra uncertainty to the calculation of weighted performance measures. Despite these limitations, we recommend using the proposed weighting to see if differences in the distribution of risk factors and verification procedures may account for discrepancies in the values of performance measures estimated in the validation population versus in the training data, rather than due to the failure of the risk model.

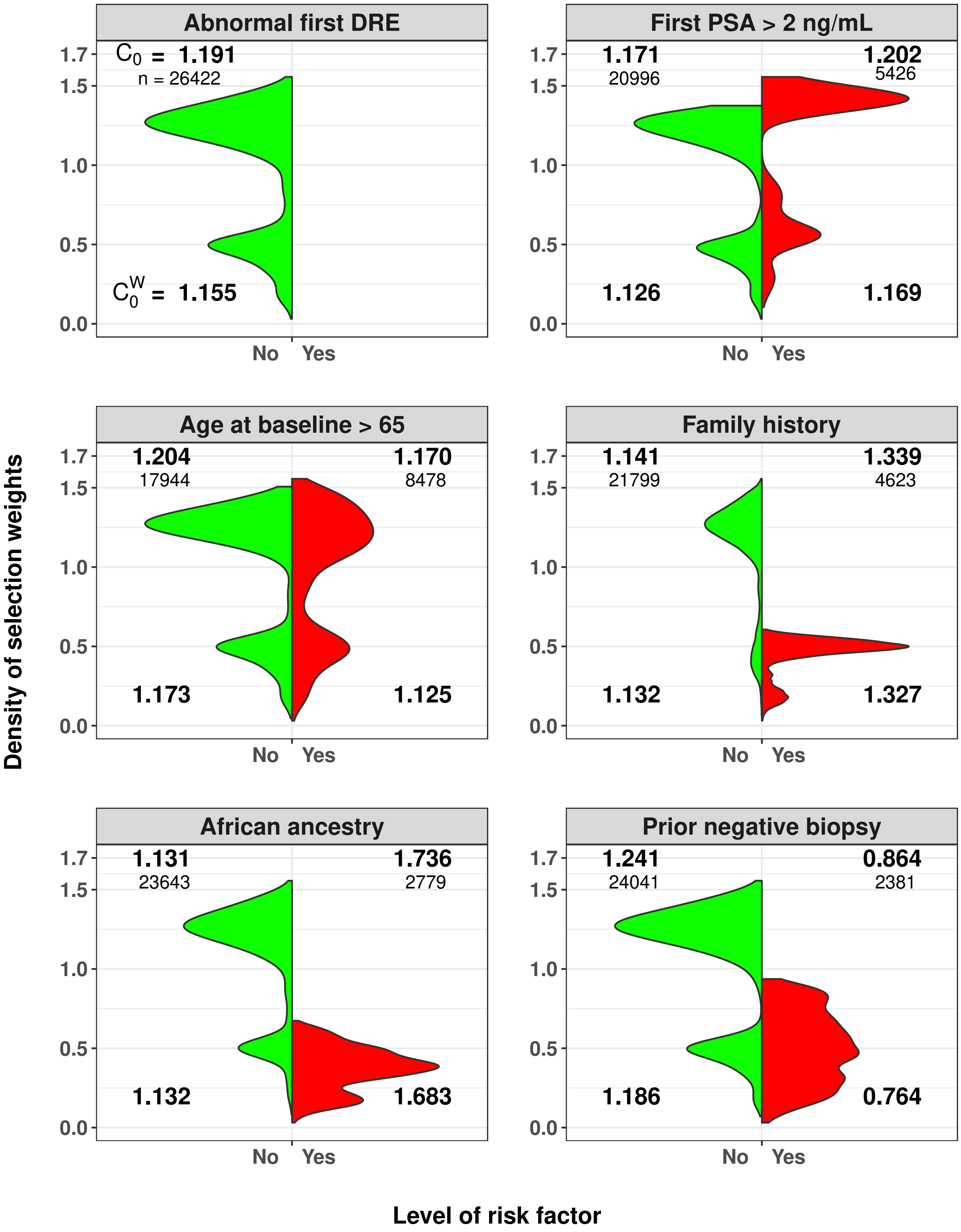

Figure 4.4: Histograms of the selection weights w from the logistic model in Table 4.7 used for calculation of $\widehat{C}_0^W$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^W (bottom numbers) calculated for each subgroup and corresponding sample sizes are shown in each panel.

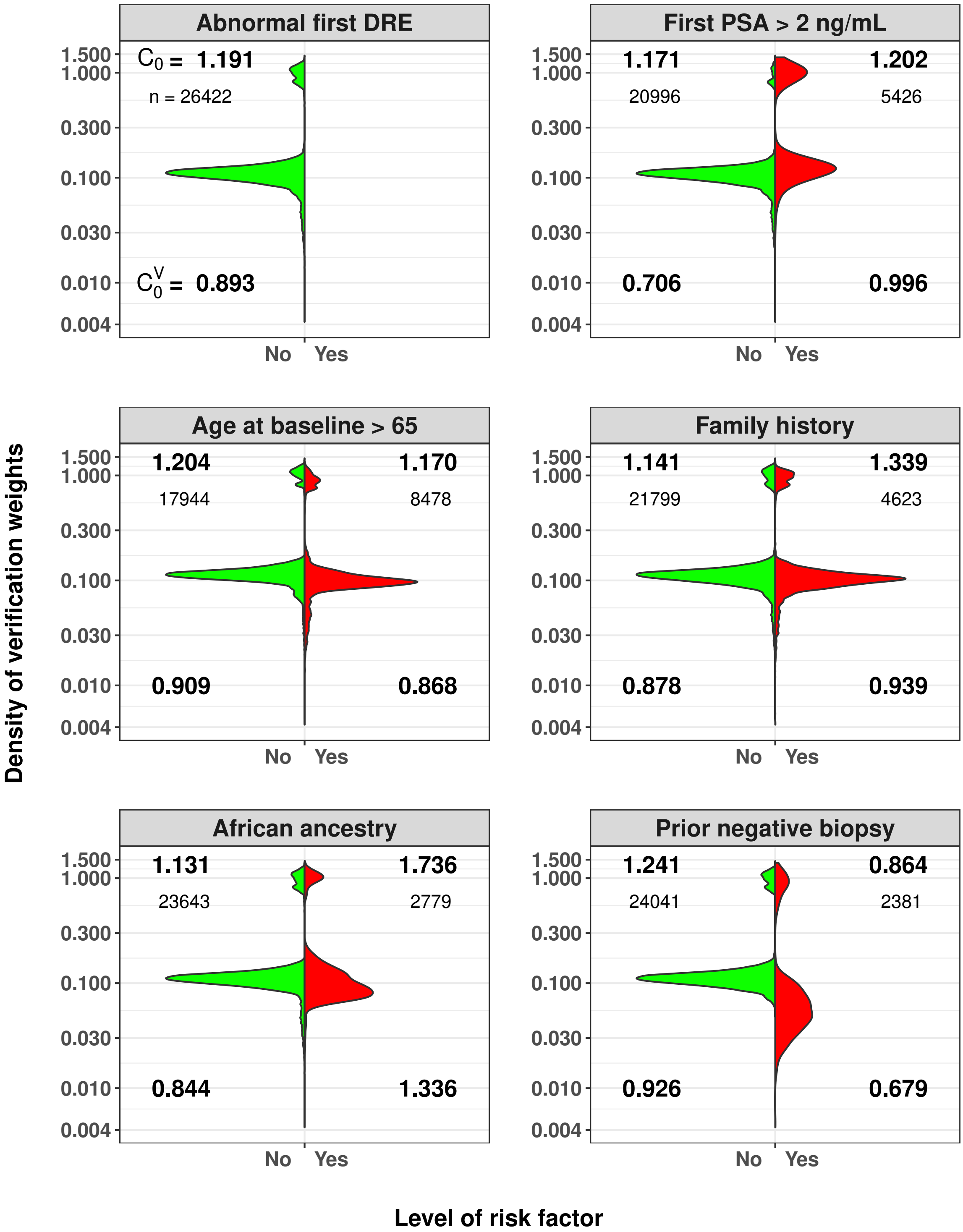

Figure 4.5: Histograms of the verification weights v from the logistic model shown in Table 4.8 used for calculation of $\widehat{C}_0^V$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^V (bottom numbers) calculated for each subgroup and corresponding sample sizes are shown in each panel. The y-axes are on the log-base-10 scale.

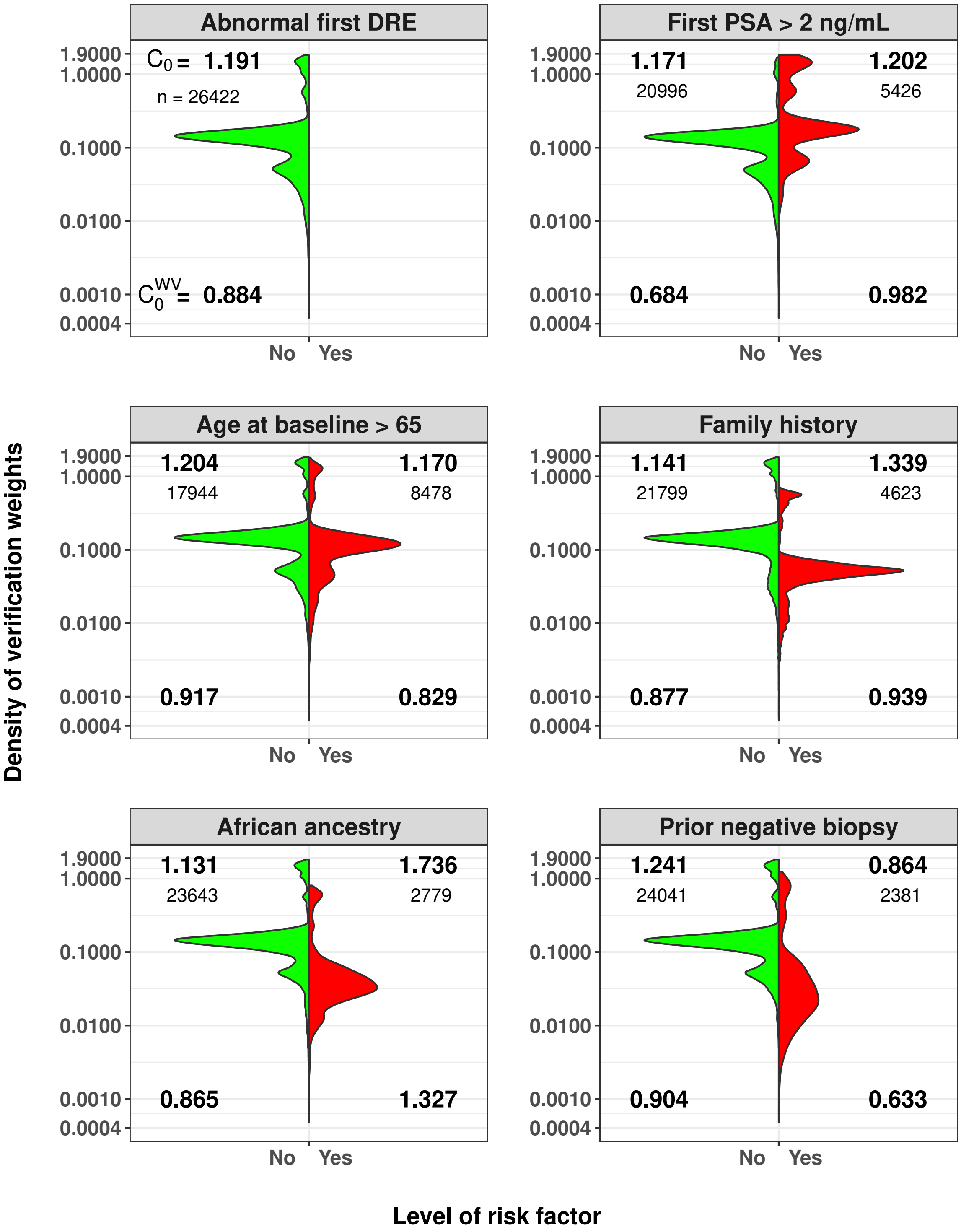

Figure 4.6: Histograms of the combined selection and verification weights used for calculation of $\widehat{C}_0^{WV}$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^{WV} (bottom numbers) calculated for each subgroup and corresponding sample sizes are shown in each panel. The selection weights w are based on the logistic model in Table 4.7 and the verification weights v are based on the logistic models in Table 4.8. The y-axes are on the log-base-10 scale.

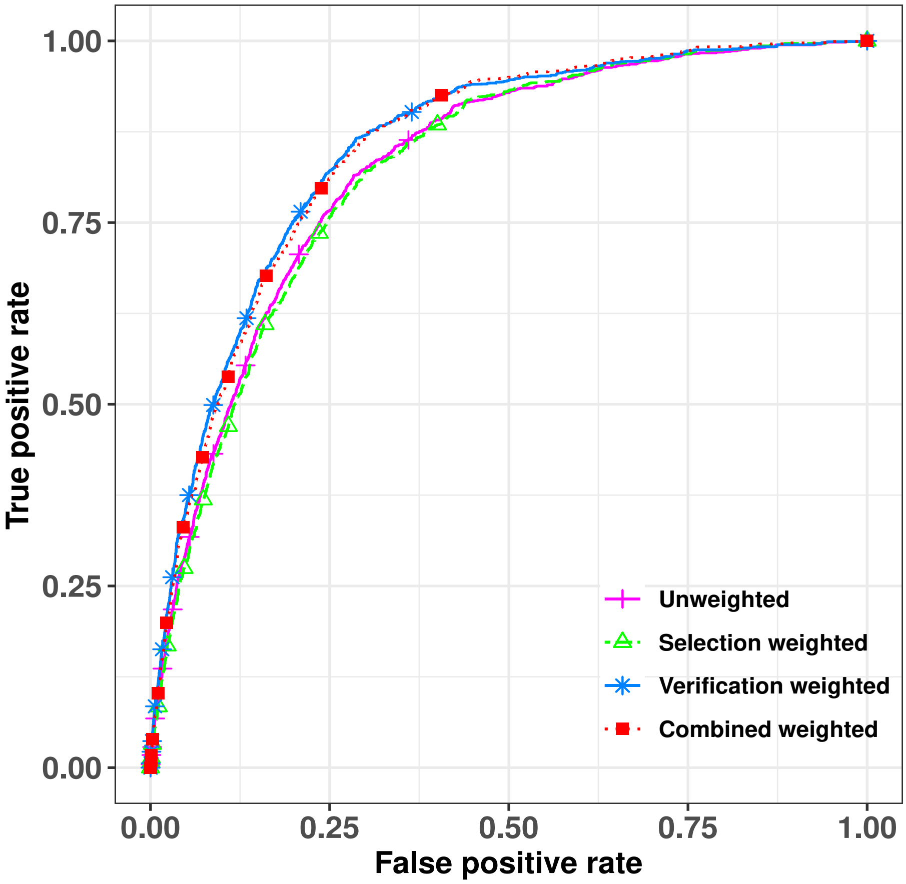

Figure 4.7: Unweighted and weighted receiver operating characteristics curves among 26422 SELECT participants. The selection weights w are based on the logistic model in Table 4.7 and the verification weights v are based on the logistic models in Table 4.8.

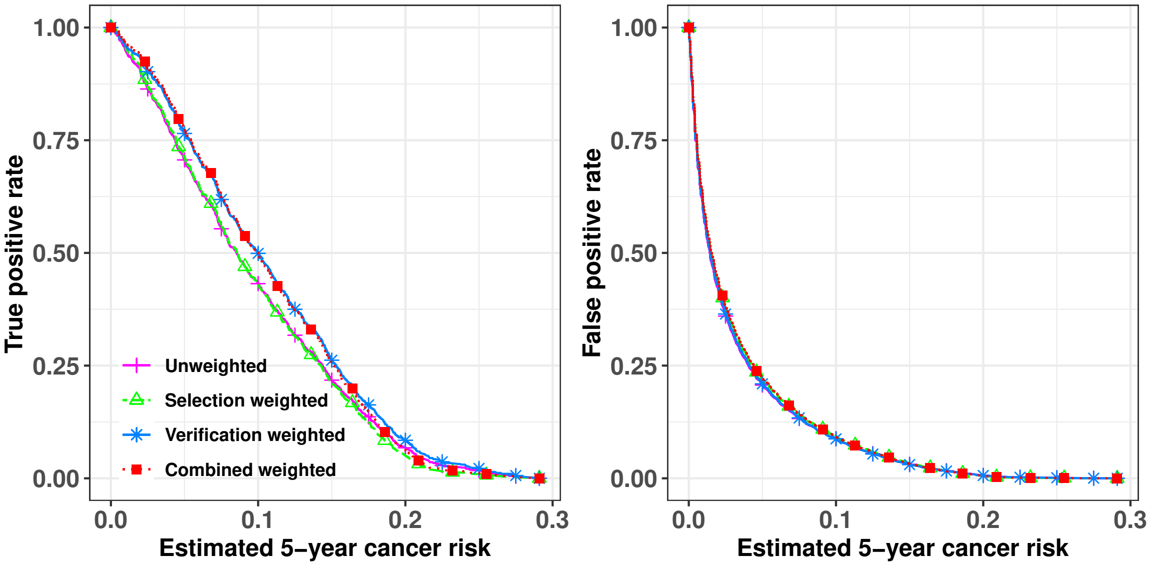

Figure 4.8: Unweighted and weighted true positive rates (TPRs) and false positive rates (FPRs) among 26422 SELECT participants. The selection weights w are based on the logistic model in Table 4.7 and the verification risks on the logistic models in Table 4.8.

Table 4.10: Odds ratios from the logistic regression model for the probability of being in PLCO without model selection with outcome (1:in PLCO versus 0: in SELECT) applied to the 56121 participants of both studies (29699 from PLCO and 26422 from SELECT).

Risk factor	Odds ratio (95% Confidence interval)
$Intercept$	2.22 (1.75, 2.82)
log_2PSA	0.75 (0.62, 0.91)
Age	0.993 (0.989, 0.996)
$Family\ history$	0.27 (0.13, 0.52)
$African\ ancestry$	0.08 (0.03, 0.21)
$Prior\ negative\ biopsy$	2.02 (0.77, 5.33)
$log_2PSA * Age$	1.006 (1.003, 1.009)
$log_2PSA * Family\ history$	1.04 (0.99, 1.10)
$log_2PSA * African\ ancestry$	1.12 (1.05, 1.20)
$log_2PSA * Prior\ negative\ biopsy$	1.17 (1.11, 1.24)
$Age * Family\ history$	1.01 (0.99, 1.02)
$Age * African\ ancestry$	1.02 (1.01, 1.04)
$Age * Prior\ negative\ biopsy$	0.98 (0.96, 0.99)
$Family\ history * African\ ancestry$	1.17 (0.92, 1.49)
$Family\ history * Prior\ negative\ biopsy$	1.13 (0.91, 1.40)
$African\ ancestry * Prior\ negative\ biopsy$	1.20 (0.86, 1.65)

Table 4.11: Characteristics table of the latest prostate-specific antigen (PSA), the indicator of the latest PSA greater than 4 ($I(PSA > 4)$), and digital rectal exam (DRE) among 29699 PLCO and 26422 SELECT participants. The latest PSA and DRE are the latest records taken before the first biopsy within 5 years for verified participants, while the latest records taken within 5 years after baseline for unverified ones.

	PLCO (N_1 = 29699)	SELECT (N_0 = 26422)
Latest PSA	(min., 1st quartile, median, 3rd quartile, max.)	
	(0.00, 0.71, 1.29, 2.50, 842.90)	(0.00, 0.70, 1.21, 2.28, 790.93)
$I(PSA > 4)$, $n(\%)$		
- 1	3909 (13.2)	2044 (7.7)
- 0	25790 (86.8)	24378 (92.3)
Latest DRE, $n(\%)$		
- Abnormal	2982 (10.0)	731 (2.8)
- Normal	26717 (90.0)	25691 (97.2)
Number of PSA tests within 5 years, $n(\%)$		
- 1	722 (2.4)	729 (2.8)
- 2	1211 (4.1)	1747 (6.6)
- 3	1473 (5.0)	2323 (8.8)
- 4	6272 (21.1)	3633 (13.7)
- 5	9230 (31.1)	9261 (35.1)
- 6	10676 (35.9)	8559 (32.4)
- 7	115 (0.4)	155 (0.6)
- 8	0 (0)	14 (0.1)
- 9	0 (0)	1 (0.0)

Table 4.12: Odds ratios estimated from logistic models for 5-year verification risk fit separately to men PLCO and SELECT. PSA=prostate-specific antigen; DRE=digital rectal exam; CI=confidence interval. For VM2 we used step-wise model selection and included up to two-way interactions of log_2PSA, $I(PSA > 4)$, DRE, age, $family\ history$, and $prior\ negative\ biopsy$. VM3 includes all factors in Table 4.8 plus the number of PSA tests. VM4 is obtained after step-wise model selection applied to a model including all main effects as for VM2, and the number of PSA tests, as well as their two-way interactions.

	PLCO (N_1 = 29699)	SELECT (N_0 = 26422)
Risk factor	Odds ratio (95% CI)	Odds ratio (95% CI)
VM2: Optimal model after model selection		
$Intercept$	0.02 (0.01, 0.03)	0.08 (0.04, 0.16)
log_2PSA	3.91 (3.33, 4.63)	4.62 (4.21, 5.09)
$I(PSA > 4)$	33.43 (24.98, 45.13)	30.70 (16.66, 55.46)
DRE	411.76 (314.99, 545.40)	88.84 (71.19, 111.12)
Age	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)
$Family\ history$	-	1.35 (1.19, 1.52)
$Prior\ negative\ biopsy$	-	1.99 (1.49, 2.64)
$log_2PSA * I(PSA > 4)$	-	0.36 (0.28, 0.47)
$log_2PSA * DRE$	0.33 (0.27, 0.39)	0.29 (0.25, 0.33)
$log_2PSA * Prior\ negative\ biopsy$	-	0.74 (0.64, 0.87)
$I(PSA > 4) * DRE$	0.08 (0.05, 0.12)	-
VM3: No model selection, including number of PSA tests		
$Intercept$	0.02 (0.01, 0.04)	0.006 (0.005, 0.008)
log_2PSA	4.67 (3.69, 5.94)	4.37 (3.97, 4.81)
$I(PSA > 4)$	390.59 (103.52, 1477.01)	24.57 (13.29, 44.52)
DRE	493.70 (352.83, 708.14)	81.83 (65.56, 102.34)
Age	0.98 (0.96, 0.99)	-
$Family\ history$	1.22 (1.03, 1.46)	1.35 (1.20, 1.52)
$African\ ancestry$	0.011 (0.001, 0.231)	-
$Prior\ negative\ biopsy$	0.79 (0.67, 0.95)	1.73 (1.30, 2.28)
$Number\ of\ PSA\ tests$	0.86 (0.83, 0.90)	1.20 (1.15, 1.25)
$log_2PSA * I(PSA > 4)$	0.74 (0.57, 0.97)	0.40 (0.31, 0.52)
$log_2PSA * DRE$	0.27 (0.21, 0.35)	0.31 (0.26, 0.37)
$log_2PSA * Prior\ negative\ biopsy$	-	0.77 (0.66, 0.90)
$I(PSA > 4) * DRE$	0.10 (0.07, 0.16)	0.64 (0.36, 1.16)
$I(PSA > 4) * Age$	0.97 (0.95, 0.99)	-
$Age * African\ ancestry$	1.07 (1.02, 1.13)	-
VM4: Optimal model after model selection including number of PSA tests as candidate risk factor		
$Intercept$	0.02 (0.01, 0.05)	0.03 (0.02, 0.07)
log_2PSA	7.12 (5.28, 9.66)	4.51 (4.11, 4.97)
$I(PSA > 4)$	109.31 (24.63, 488.46)	26.79 (14.48, 48.53)
DRE	127.42 (72.89, 226.86)	84.10 (67.35, 105.25)

Age	0.98 (0.97, 1.00)	0.97 (0.96, 0.98)
$Family\ history$	-	1.33 (1.18, 1.50)
$Prior\ negative\ biopsy$	-	1.95 (1.46, 2.58)
$Number\ of\ PSA\ tests$	0.82 (0.74, 0.92)	1.20 (1.15, 1.25)
$log_2 PSA * I(PSA > 4)$	-	0.38 (0.29, 0.49)
$log_2 PSA * DRE$	0.35 (0.29, 0.42)	0.29 (0.25, 0.34)
$log_2 PSA * Prior\ negative\ biopsy$	-	0.75 (0.64, 0.88)
$log_2 PSA * Number\ of\ PSA\ tests$	0.87 (0.82, 0.92)	-
$I(PSA > 4) * DRE$	0.08 (0.05, 0.12)	-
$I(PSA > 4) * Age$	0.96 (0.94, 0.98)	-
$I(PSA > 4) * Number\ of\ PSA\ tests$	1.35 (1.18, 1.55)	-
$DRE * Number\ of\ PSA\ tests$	1.26 (1.13, 1.41)	-

Table 4.13: Sensitivity analysis. Estimated calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants using D in the calculation of the measures. 95% confidence intervals (CIs) are based on percentiles of the bootstrap empirical distribution function with 600 bootstrap repetitions. The model for w including all risk factors and their two-way interactions without model selection (MS) is shown in Table 4.10. The model for v without MS is shown in Table 4.8. Model VM2 for v is the optimal model after step-wise MS allowing up to two-way interactions of log_2PSA, $I(PSA > 4)$, DRE, Age, $Family\ history$, and $Prior\ negative\ biopsy$. Model VM3 for v includes all factors in Table 4.8 plus the number of PSA tests. Model VM4 for v is the optimal model after step-wise model selection applied to to a model including the same base model as for VM2 plus the number of PSA tests, and their two-way interactions. Coefficients for v from VM2, VM3, and VM4 are shown in Table 4.12.

	Estimate	95% CI	Model for weights
C_0^W	1.157	(1.088, 1.226)	w, no MS
C_0^V	0.878	(0.826, 0.942)	v, VM2, MS
C_0^V	0.858	(0.803, 0.918)	v, VM3, no MS
C_0^V	0.833	(0.782, 0.900)	v, VM4, MS
C_0^{WV}	0.875	(0.819, 0.937)	v, VM2, MS; w, MS
C_0^{WV}	0.849	(0.789, 0.908)	v, VM3, no MS; w, MS
C_0^{WV}	0.828	(0.775, 0.893)	v, VM4, MS; w, MS
C_0^{WV}	0.886	(0.826, 0.944)	v, no MS; w, no MS
C_0^{WV}	0.877	(0.820, 0.939)	v, VM2, MS; w, no MS
C_0^{WV}	0.851	(0.790, 0.912)	v, VM3, no MS; w, no MS
C_0^{WV}	0.830	(0.778, 0.897)	v, VM4, MS; w, no MS
AUC_0^W	0.825	(0.813, 0.836)	w, no MS
AUC_0^V	0.853	(0.841, 0.865)	v, VM2, MS
AUC_0^V	0.852	(0.840, 0.864)	v, VM3, no MS
AUC_0^V	0.852	(0.839, 0.864)	v, VM4, MS
AUC_0^{WV}	0.851	(0.839, 0.861)	v, VM2, MS; w, MS
AUC_0^{WV}	0.850	(0.837, 0.861)	v, VM3, no MS; w, MS
AUC_0^{WV}	0.850	(0.837, 0.861)	v, VM4, MS; w, MS
AUC_0^{WV}	0.851	(0.839, 0.862)	v, no MS; w, no MS
AUC_0^{WV}	0.851	(0.840, 0.862)	v, VM2, MS; w, no MS
AUC_0^{WV}	0.850	(0.838, 0.862)	v, VM3, no MS; w, no MS
AUC_0^{WV}	0.850	(0.838, 0.861)	v, VM4, MS; w, no MS

Table 4.14: Sensitivity analysis. Estimated calibration ratios (C_0, C_0^W, C_0^V, C_0^{WV}) and areas under the receiver operating characteristic curves (AUC_0, AUC_0^W, AUC_0^V, AUC_0^{WV}) in 26422 SELECT participants using the censoring weighted outcome, D^C, in the calculation of the measures. 95% confidence intervals (CIs) are based on percentiles of the bootstrap empirical distribution function with 600 bootstrap repetitions. The model for w including all risk factors and their two-way interactions without model selection (MS) is shown in Table 4.10. The model for v without MS is shown in Table 4.8. Model VM2 for v is the optimal model after step-wise MS including up to two-way interactions of log_2PSA, $I(PSA > 4)$, DRE, Age, $Family\ history$, and $Prior\ negative\ biopsy$. Model VM3 for v includes all factors in Table 4.8 plus the number of PSA tests. Model VM4 for v is the optimal model after step-wise model selection applied to a model including the same main effects as in the base model of VM2 plus the number of PSA tests, and their two-way interactions. Coefficients for v from VM2, VM3, and VM4 are shown in Table 4.12.

	Estimate	95% CI	Model for weights
C_0	1.209	(1.142, 1.277)	–
C_0^W	1.172	(1.101, 1.239)	w, MS
C_0^W	1.174	(1.104, 1.244)	w, no MS
C_0^V	0.894	(0.840, 0.953)	v, no MS
C_0^V	0.879	(0.827, 0.943)	v, VM2, MS
C_0^V	0.859	(0.804, 0.919)	v, VM3, no MS
C_0^V	0.834	(0.784, 0.901)	v, VM4, MS
C_0^{WV}	0.885	(0.825, 0.942)	v no MS; w, MS
C_0^{WV}	0.876	(0.820, 0.938)	v, VM2, MS; w,MS
C_0^{WV}	0.850	(0.790, 0.909)	v, VM3, no MS; w, MS
C_0^{WV}	0.829	(0.776, 0.894)	v, VM4, MS; w, MS
C_0^{WV}	0.887	(0.827, 0.945)	v, no MS; w, no MS
C_0^{WV}	0.878	(0.821, 0.940)	v, VM2, MS; w no MS
C_0^{WV}	0.852	(0.791, 0.913)	v, VM3, no MS; w no MS
C_0^{WV}	0.831	(0.778, 0.897)	v, VM4, MS; w no MS
AUC_0	0.826	(0.814, 0.838)	–
AUC_0^W	0.822	(0.810, 0.834)	w, MS
AUC_0^W	0.823	(0.811, 0.834)	w, no MS
AUC_0^V	0.853	(0.842, 0.865)	v, no MS
AUC_0^V	0.853	(0.841, 0.865)	v, VM2, MS
AUC_0^V	0.852	(0.840, 0.864)	v, VM3, no MS
AUC_0^V	0.852	(0.839, 0.864)	v, VM4, MS
AUC_0^{WV}	0.851	(0.839, 0.862)	v, no MS; w, MS
AUC_0^{WV}	0.851	(0.839, 0.861)	v, VM2, MS; w, MS
AUC_0^{WV}	0.850	(0.837, 0.861)	v, VM3, no MS; w, MS
AUC_0^{WV}	0.850	(0.837, 0.861)	v, VM4, MS; w, MS
AUC_0^{WV}	0.851	(0.839, 0.862)	v, no MS; w, no MS
AUC_0^{WV}	0.851	(0.840, 0.862)	v, VM2, MS; w, no MS
AUC_0^{WV}	0.850	(0.838, 0.862)	v, VM3, no MS; w, no MS
AUC_0^{WV}	0.850	(0.838, 0.861)	v, VM4, MS; w no MS

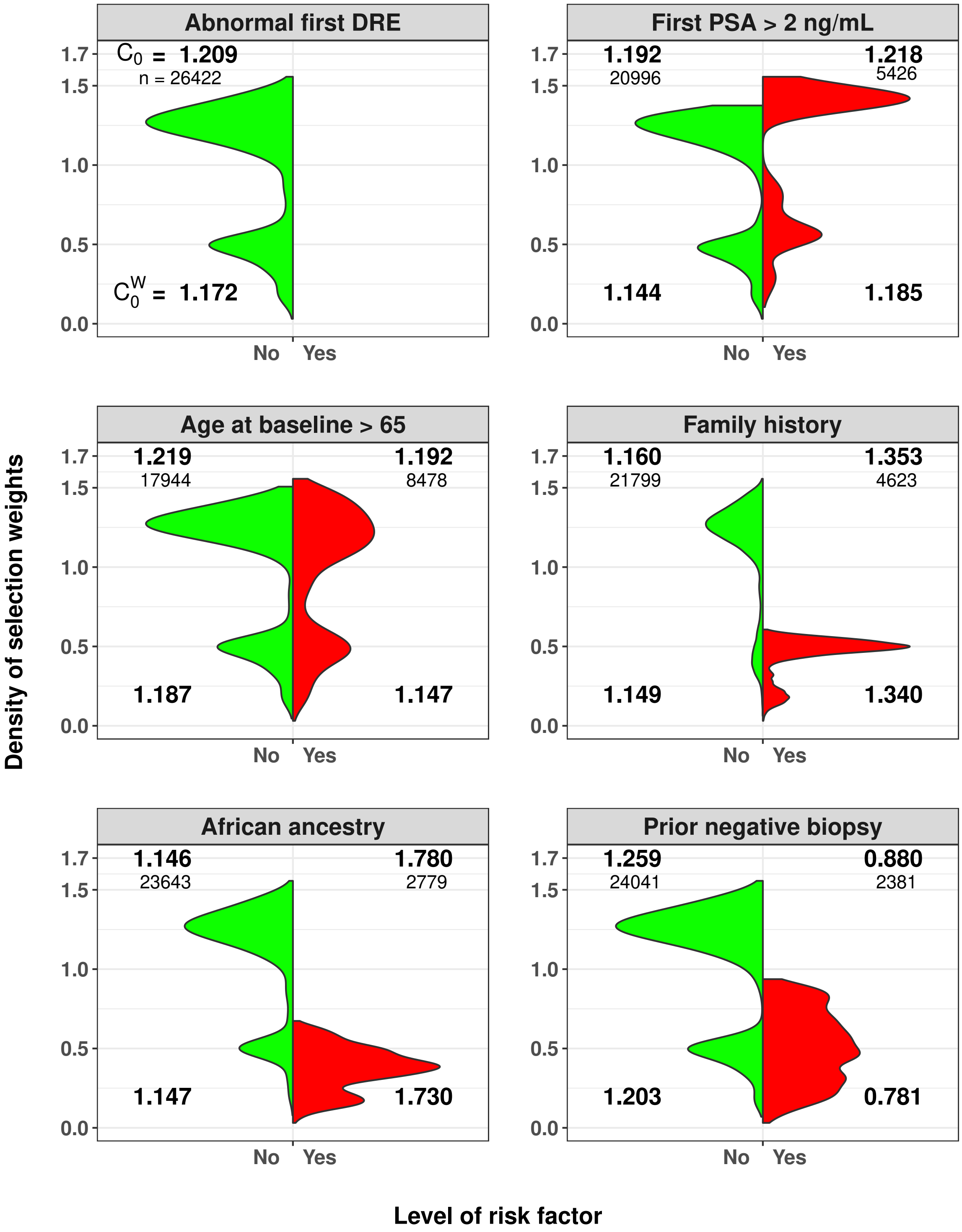

Figure 4.9: Histograms of the selection weights w from the logistic model in Table 4.7 used for calculation of $\widehat{C}_0^W$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^W (bottom numbers) are calculated using D^C, censoring weighted outcome. For each subgroup, the corresponding sample sizes are shown in each panel.

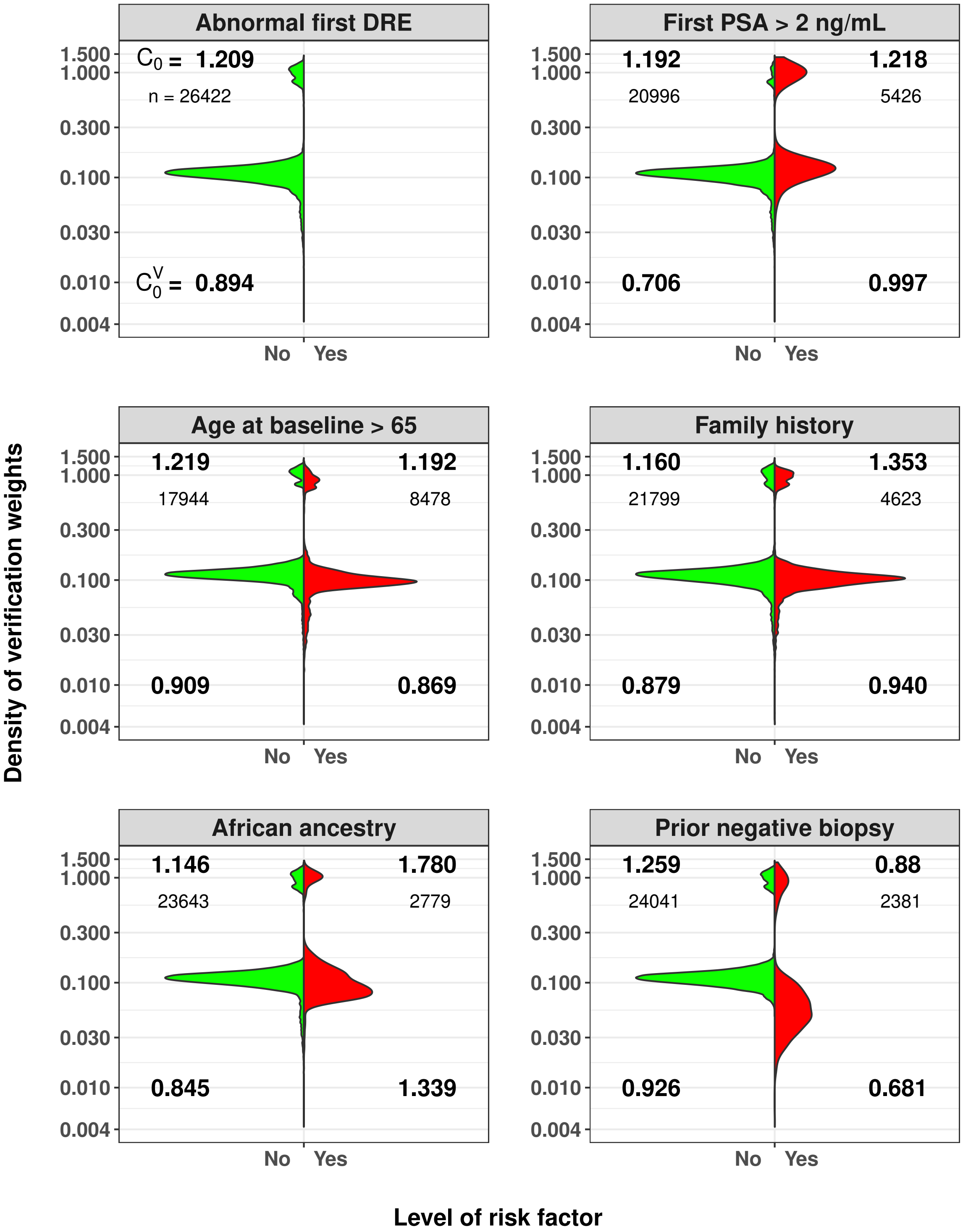

Figure 4.10: Histograms of the verification weights v from the logistic model shown in Table 4.8 used for calculation of $\widehat{C}_0^V$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^V (bottom numbers) are calculated using D^C, censoring weighted outcome. For each subgroup, the corresponding sample sizes are shown in each panel. The y-axes are on the log-base-10 scale.

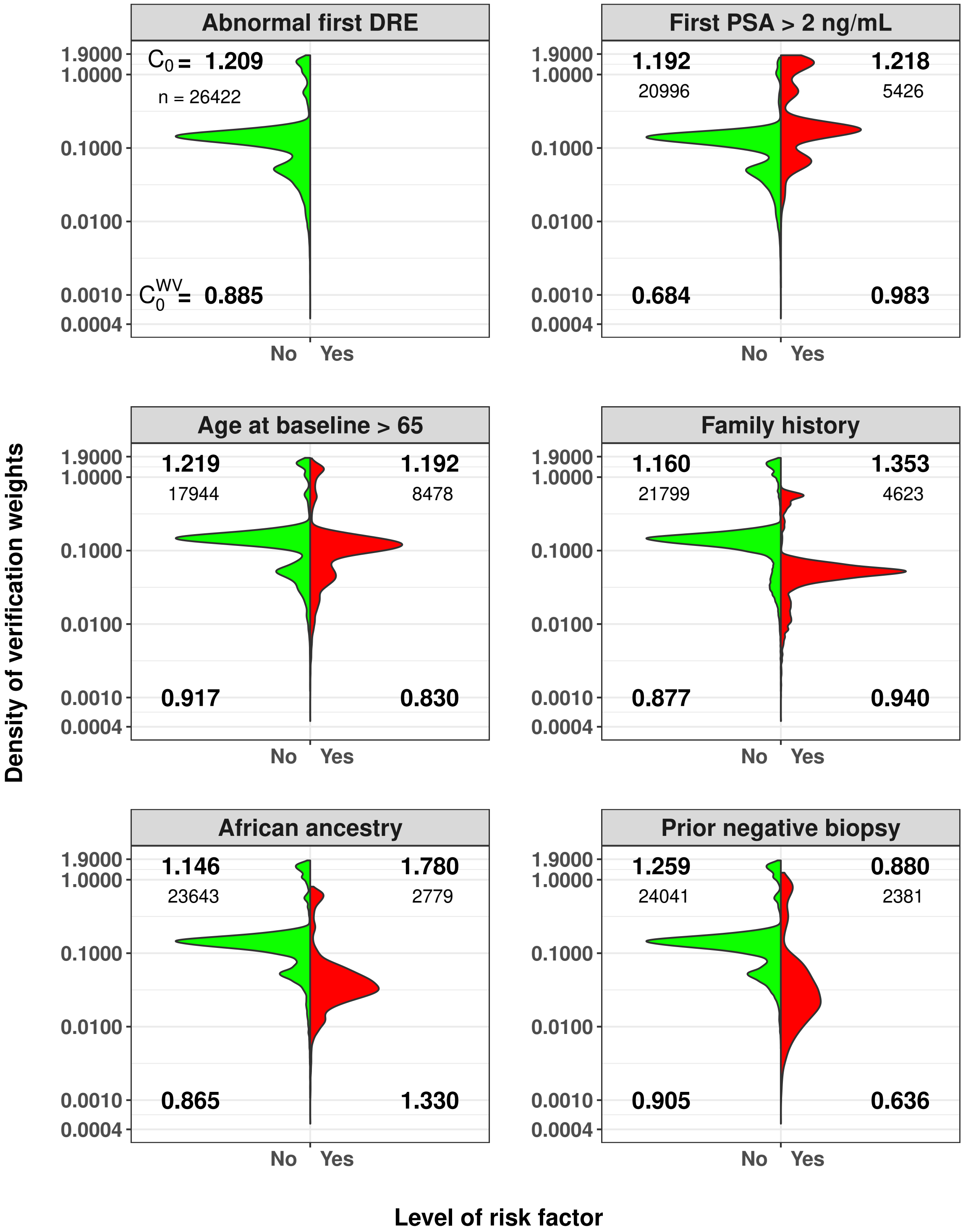

Figure 4.11: Histograms of the combined selection and verification weights used for calculation of $\widehat{C}_0^{WV}$ for the 26422 SELECT participants according to baseline risk factor categories. C_0 (top numbers) and C_0^{WV} (bottom numbers) are calculated using D^C, censoring weighted outcome. For each subgroup and corresponding sample sizes are shown in each panel. The selection weights w are based on the logistic model in Table 4.7 of the main manuscript and the verification weights v are based on the logistic model in Table 4.8 of the main manuscript. The y-axes are on the log-base-10 scale.

5 Summary and outlook

Nowadays, many clinical risk prediction models are available online assisting people worldwide in the early detection, diagnosis, or prognosis of various types of diseases based on the respective risk factors. Though aiming for different diseases, these tools all give the estimated probability of having the event of interest, i.e. the risk of experiencing the event, given the observed values of risk factors. Once a risk prediction model has been built, its performance should be at least validated internally within the training population, which is known as internal validation and is often known as assessing the “reproducibility” of the model since it assesses the performance of the model upon the same population as the underlying one used to build this model [Steyerberg, 2019]. One can validate a prediction model internally by simply applying it to the training data or following a cross-validation process [Steyerberg, 2019]. In the cross-validation process, we would first split the training data into multiple samples. We then use each of these samples as the validation set to assess the refitted model that has been re-trained on the rest of the samples.

Internally validated may not be the ultimate goal of developing risk prediction tools that are planned to be released online as people around the world would apply these tools under various circumstances to gain insights into the risk of events, such as applying them to populations from different countries with a different ethnicity than the underlying ones used for training those models. In such case, the objected population for applying a built model may have different distributions of the risk factors than the training one, which differences could attribute to geographical or temporal variations or usage of different clinical trial designs between the training sample and the objected one. Only performing well internally is not enough. One should evaluate the performance of the model with external samples that have not been used to build the model, namely external validation, which is used to assess the “transportability” of the model [Steyerberg, 2019]. Due to the potential differences in the population characteristics, which are mainly presented as having different distributions of risk factors, the external validation results may bias. We may draw wrong conclusions and falsely recommend extending the application of the model that is just valid in certain kinds of populations based on such distorted external validation results.

For both internal and external validation, the calibration and the discrimination of the developed model are two aspects to be assessed. The calibration of the model refers to the degree that which the predictions from the model agree with the observed outcomes. The calibration-in-the-large (CIL) and the calibration ratio are the two common measures quantifying the calibration of a model, where the former is based on the difference between predictions and observations with an ideal value of zero and the latter is based on the ratio of them with ideal value one. We can also visualize the calibration of a model via a calibration plot sketching observed risk on the y-axis versus predicted value on the x-axis. Participants are grouped before plotting, usually by the percentiles of predicted probabilities, to compute the observed prevalence of outcomes within the group. The discrimination of a model reflects if it can accurately predict and correctly distinguish the participants with versus without the event of interests such that the model gives higher risk values to those participants with observed events, while lower for the rest. The visualization of discrimination of a model is by the receiver operating characteristic curve (ROC) showing the true positive rates on the y-axis versus the false positive rates on the x-axis at different threshold values. A ROC bending towards the top left is desirable. If the outcome status is binary, a quantitative measure of discrimination related to the ROC is the area under the ROC (AUC) with a value closer to one indicating a better ability of discrimination. Other than AUC, we can also use the true positive rate and the false positive rate [Pearce and Ferrier, 2000; Jehi et al., 2020], the mean difference of the estimated predictions between the groups with and without the event of interest [Steyerberg et al., 2010; Pencina and D'Agostino, 2015], or the plot with the true and false positive rate on the y-axis versus the threshold for assessing the discrimination of the model in practice [Pepe and Janes, 2013].

The heterogeneous between the training and the external validation cohorts in the distributions of risk factors, also known as "case-mix difference" [Steyerberg, 2019], can distort the external validation results leading to fallacious conclusions about the validity of the developed model upon external population. When the individual participant data from both training and validation cohorts are available, where training data are often not publicly available or available only upon required, we can directly compare the distributions of risk factors between cohorts. The characteristics tables or distribution plots are commonly used for such comparison [Ankerst et al., 2018; Tolksdorf et al., 2019]. Debray et al. [2015] used logistic regression with the cohort indicator as the response developed on the data pooling two cohorts together to check if the two cohorts are similar, where two cohorts may be more similar if the internally estimated AUC of this logistic model is away from one. Powers et al. [2019] weighted the validation sample to obtain a cohort resembling the target population. Given that the developed prediction model would be perfect for the validation cohort if the training and validation cohorts are from the same population with the same distributions of risk factors, Vergouwe et al. [2010] proposed to and later, applied by Austin et al. [2016], simulate the outcomes for the validation cohort based on the predictions and then, calculated the performance measures using the simulated outcomes. Such values would be free of the impact of the heterogeneity in the distributions of features between training and validation populations. van Klaveren et al. [2016] replaced the comparison of the observed outcomes with the comparison of the estimated linear predictors in the calculation of AUC with external data, obtaining an AUC value that one would have if the prediction model is correct for the validation set, similarly to the simulation-based values used in Vergouwe et al. [2010]. These benchmark values in Vergouwe et al. [2010] and Debray et al. [2015] do not require the availability of the training data.

From the perspective of matching or balancing the distributions of risk factors between two cohorts, one can find a rich literature using propensity score weighting/matching in various research areas like survey research and causal inference [Rosenbaum and Rubin, 1983; Dahabreh et al., 2019; Kern et al., 2016; Westreich et al., 2017; Elliot, 2013; Schonlau et al., 2017; Elliott et al., 2017], where the propensity scores are the probability of being in one cohort versus the others from a logistic model built on the data pooling both cohorts together with the cohort indicator as the response. Ackerman et al. [2019] used the propensity score weighting resembling the validation cohort to the training one, where participants in the validation who represents the training population better would have higher weights.

In the development of the risk prediction model, not all participants would be used to train the model but only those with known status of the event of interest. Often, the participants will go through a verification process to ascertain their outcomes, like receiving a prostate biopsy to confirm the status of prostate cancer. Some verification procedures are invasive, such as the biopsy to ascertain cancer that needs surgery to sample tissue from the organ. To avoid unnecessary verification and wasting of resources, the participants who are sent to verify should meet certain criteria. For example, in the prostate cancer screening trial, only men with PSA levels above the threshold, usually 4 ng/mL, or having abnormal DRE would be recommended to receive a biopsy. In the COVID-19 pandemic, due to limited laboratory load, usually, only people who are in close contact or have positive antigen rapid test results would be recommended to have the polymerase chain reaction (PCR) test to confirm the status of infection. Without the verification process, the outcome status of a participant is unknown and hence, missing. Because there are criteria of recommending to be verified, even within the same cohort, the characteristics of participants who have been verified may differ from the unverified ones and hence, verification bias appears. The verified participants are not a random sample from the population when the bias occurs. To address the verification bias in the calculation of the true and false positive rates for the assessment of the accuracy of a prediction model, Begg and Greenes [1983] and Alonzo and Pepe [2005] imputed the missing outcome for the unverified participants with the estimated risk from the model built with the verified participants data under the missing-at-random (MAR) assumption. In the external validation, the verification mechanisms, such as the criteria for recommending to be biopsied, could differ between cohorts due to different trial designs, where the verification bias between cohorts may exist and affect the external validation results as well. We checked their proposed verification bias adjustment method under MAR assumption in Chapter 3.

In this thesis, we focus on addressing the impact of the differences in the distribution of risk factors and verification process between the training and validation populations in the external validation of the clinical risk prediction model, given that the data from both populations are available. In Chapter 4, we introduced a novel weighting framework to address the selection and verification bias, resembling the validation population to the training one in the risk factors distribution and disease prevalence. In this setting, we first presumed that the true disease probability depends on the risk factors ($\mathbf{X}^*$) consisting of the predictors used in the prediction model ($\mathbf{X}$) as well as the omitted features ($\mathbf{Z}$). We also assumed that the developed risk prediction model ($R(\mathbf{X})$) is well calibrated and accurate with the internal calibration ratio and AUC both equal to one.

When no verification bias occurs but only selection bias, we checked the performance of the selection weighted calibration ratio and AUC against the corresponding internally estimated unweighted values and formalized the idea of "reproducibility" and "transportability". When both populations have equal true disease risk and the same conditional distribution of the omitted features ($\mathbf{Z}$) given predictors ($\mathbf{X}$), the unweighted and selection weighted calibration ratios and AUCs on the validation cohort are equal to the respective unweighted values estimated with training data if additionally assuming the distributions of predictors are the same between populations ($F_0(\mathbf{X}) = F_1(\mathbf{X})$). In this case, the model is reproducible in the validation population since it is well calibrated and discriminated. External validation with such a population is just the same as internal validation. On the other hand, if the additional assumption that $F_0(\mathbf{X}) = F_1(\mathbf{X})$ does not hold, only the selection weighted calibration ratio and AUC on the validation cohort are ideally one, but not their corresponding unweighted measures. The model is transportable, i.e. the unweighted calibration ratio on validation cohort is one, only if the risk prediction model ($R(\mathbf{X})$) is strongly calibrated in the training data in this case. Here, the strongly calibrated refers to the true disease risk given $\mathbf{X}$ is the same as the prediction $R(\mathbf{X})$ for any $\mathbf{X}$ values. When validation bias occurs, instead of modeling the risk of disease, we rather model the risk of having the disease as well as being verified. We weight the observations, i.e. the numerator of the calibration ratio in our expression, but not the prediction to adjust the verification bias. By assuming that the training and validation populations have the same distribution of $\mathbf{X}^*$ and the same conditional probability of having the disease given $\mathbf{X}^*$ and be verified, one could have the verification weighted, but not the unweighted, calibration ratio be one.

From the simulation study, we showed that if no verification bias occurs but only selection bias ($S2, V1$ and $S3, V1$), the selection weighted or combined selection and verification weighted measures are less biased than the unweighted or verification weighted ones given that the model for selection bias is correctly specified. When both risk factor distribution and the verification process differ between the populations, only the combined weighted measures could substantially reduce the bias. In the application with PLCO and SELECT, we studied the 5-year prostate cancer risk via the Cox regression using age as the time scale. As shown in Table 4.9, the weighted values improved compared to the unweighted one regardless of the types of weighting except that the selection weighted AUC was slightly lower than the unweighted one. We recommend using the proposed weighting methods to accommodate the selection and verification biases, checking if the poor model performance on the validation population may be rather due to the bias between populations than the failure of the developed model.

There are some limitations to the proposed methods. The weighting framework requires the availability of both training and validation data to be used for modeling the weights. We have to correctly specify the models for the weights to properly adjust the bias, which is a challenge in practice because trials have different designs and we could miss crucial risk factors in the models for bias adjustment. For the adjustment for selection bias, we need the same risk factors from both cohorts. The number of risk factors used for adjusting selection bias may be very limited because trials can have different assessment plans and data sharing policies such that not all features available in one trial would also be available in the other. Even if the same risk factors have been collected in both training and validation data, the underlying collection methods could be different, such as using different devices that the modern ones may give more precise results compared to the previous ones. Because all proposed performance measures rely on weights, the variation of the resulting performance measures may increase due to different choices of weighting models. Though with limitations, we still recommend applying the proposed bias-adjusted measures to have an insight into the impact of selection bias or verification bias or both on the external validation results, where the modeling of weights should be carefully considered.

We can extend the proposed framework to adjust other biases occurring in the external validation prediction model. Due to advancing in the medical apparatus techniques and the modernization of disease diagnosis and prevention standards, prediction tools developed with past data may not be valid for samples from later time points. Therefore, the measurements of risk factors may inevitably systematically differ, resulting in different predictor effects between samples or the underlying diagnosis process differ such that the timing of disease detection could regularly vary. In this case, the temporal bias could occur in the validation of the tools with the latest samples. In the COVID-19 pandemic, the rapidly changing screening and reporting policies and variations of the virus demand constant calibration of the developed prediction tools. One may not able to gain enough qualified samples suitable for assuring the external validity of these tools because of the poor data reporting quality, different referral policies across countries, let alone the heterogeneous medical history of participants and different pandemic phases the participants at. When using the samples at hand to validate the developed COVID-19 risk prediction tools, we should consider adjusting for the explicit bias between the validation and training data. For example, if a potential COVID-19 infected participant is first screened at the primary care site and then, referred to secondary medical care for further diagnosis, we can introduce the referral weights that may consist of the probability of referral, in addition to verification weights, to accommodate the referral bias under our bias adjustment framework. We can define the weights used for bias adjustment flexibly depending on the reality and the need. To conclude, one should have an omnibus view of the training and validation populations, including the knowledge of the designs of trials, variable collection methods, and diagnosis criteria, to properly adjust the bias in the external validation of the developed prediction model. We encourage researchers to disclose relevant information and share the underlying data when publishing risk prediction models to facilitate external validation.

An alternative method to address the bias in the external validation performance assessments due to heterogeneity in the distributions of risk factors between the training and validation cohorts would be from a causal inference perspective using potential outcomes [Neyman, 1923; Rubin, 1974]. Specifically, current definitions of the complier average causal effect (CACE) and survival average causal effect (SACE) could be extended to the define a concept of verified average causal effect (VACE) or selected and verified average causal effect (SVACE) in external validation studies [Guo et al., 2022; Hayden et al., 2005].

CACE is routinely used for causal inference to determine treatment effects in randomized trials among the principal stratum of always-compliers, defined as patients who would have complied with treatment regardless of randomized assignment. The principal stratum of always-compliers is not identifiable, and thus requires assumptions and sensitivity analyses to violations of the assumptions. With $Z = 0, 1$ denoting randomized treatment assignment, the hypothetical compliance indicators $C(0), C(1)$ under each treatment arm are modeled for each participant, as well as the hypothetical outcomes $O(0), O(1)$ on each treatment arm. Each participant thus has four potential outcomes, $C(0), C(1), O(0), O(1)$ for which only half are observed, namely those for the treatment arm the participant was observed to be assigned. Conditional on covariates X, CACE was defined as

$$CACE = E(O(1) - O(0)|C(0) = C(1) = 1, X).$$

SACE is defined similarly as the treatment effect on outcomes measurable only for survivors for the always-survivors, defined as patients who would have survived under both treatment arms.

The potential outcomes framework can be extended to the selection and verification processes depicted in Figure 3.1, by assigning $Z = 1, 0$ to indicate the training and validation set, $T(0), T(1)$ patient selection indicators into each set, $V(0), V(1)$ verification indicators for each set, and X the baseline risk factors, PSA, DRE, and so forth. Then instead of the difference in outcomes $O(1) - O(0)$ as the primary endpoint, SVACE would use $R(X) - D$, which is a risk model and is the outcome of prostate cancer evaluated on the principal stratum,

$$SVACE = E(R(X) - D|T(0) = T(1) = V(0) = V(1) = 1, X).$$

We are currently working on identifying the minimal set of assumptions needed to identify SVACE. From the parallels to CACE and SACE, the resulting estimators should be weighted sums as in (3.9), for which the same consistency theorems would hold.

Contributing manuscript

Chapter 4 in this thesis has been accepted for publication in the journal Statistics in Medicine.

*Pfeiffer, Ruth M., **Chen, Yiyao**, Gail, Mitchell H., Ankerst, Donna P. (2022). Accommodating population differences in model validation. (accepted)*

Contributions of authors: Yiyao Chen performed all statistical analyses and simulation, created the tables and figures, drafted the manuscript, and critically revised the manuscript. Ruth M. Pfeiffer and Donna P. Ankerst supervised the study, assisted in structuring the manuscript, made limited contribution to manuscript writing, assisted in revision, and handled the review process. Mitchell H. Gail contributed to the revision and editing.

Acronyms

AUC	Area-under-the-receiver-operating-characteristic curve
CI	Confidence interval
CIL	Calibration-in-the-large
DRE	Digital rectal examination
ERSPC	European Randomized Study of Screening for Prostate Cancer
LP	Linear predictor
MAR	Missing-at-random
ng/mL	Nanogram per milliliter
PBCG	Prostate Biopsy Collaborative Group
PBCG-RC	Prostate Biopsy Collaborative Group Risk Calculator
PLCO	Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial
PCA	Prostate cancer
PSA	Prostate-specific antigen
ROC	Receiver operating characteristic curve
SELECT	Selenium and Vitamin E Cancer Prevention Trial

Bibliography

Ackerman, B, Siddique, J, and Stuart, EA, (2019). Transportability of outcome measurement error correction: from validation studies to intervention trials. *arXiv preprint arXiv:1907.10722*.

Adimari, G and Chiogna, M, (2015). Nearest-neighbor estimation for ROC analysis under verification bias. *The International Journal of Biostatistics*, 11(1):109–124.

Adimari, G and Chiogna, M, (2017). Nonparametric verification bias-corrected inference for the area under the roc curve of a continuous-scale diagnostic test. *Statistics and Its Interface*, 10(4):629–641.

Alberts, AR, Roobol, MJ, Verbeek, JF, Schoots, IG, Chiu, PK, Osses, DF, Tijsterman, JD, Beerlage, HP, Mannaerts, CK, Schimmöller, L, et al., (2019). Prediction of high-grade prostate cancer following multiparametric magnetic resonance imaging: improving the Rotterdam European randomized study of screening for prostate cancer risk calculators. *European Urology*, 75(2):310–318.

Alonzo, TA and Pepe, MS, (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):173–190.

Andriole, GL, Crawford, ED, Grubb III, RL, Buys, SS, Chia, D, Church, TR, Fouad, MN, Gelmann, EP, Kvale, PA, Reding, DJ, et al., (2009). Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319.

Andriole, GL, Crawford, ED, Grubb III, RL, Buys, SS, Chia, D, Church, TR, Fouad, MN, Isaacs, C, Kvale, PA, Reding, DJ, Weissfeld, JL, Yokochi, LA, O'Brien, B, Ragard, LR, Clapp, JD, Rathmell, JM, Riley, TL, Hsing, AW, Izmirlian, G, Pinsky, PF, Kramer, BS, Miller, AB, Gohagan, JK, Prorok, PC, and PLCO Screening Trial Project Team, (2012). Prostate cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial: mortality results after 13 years of follow-up. *Journal of the National Cancer Institute*, 104(2):125–132.

Ankerst, DP, Boeck, A, Freedland, SJ, Thompson, IM, Cronin, AM, Roobol, MJ, Hugosson, J, Jones, JS, Kattan, MW, Klein, EA, et al., (2012). Evaluating the PCPT risk calculator in ten international biopsy cohorts: Results from the Prostate Biopsy Collaborative Group. *World Journal of Urology*, 30(2):181–187.

Ankerst, DP, Hoefler, J, Bock, S, Goodman, PJ, Vickers, A, Hernandez, J, Sokoll, LJ, Sanda, MG, Wei, JT, Leach, RJ, and Thompson, IM, (2014). Prostate cancer prevention trial risk calculator 2.0 for the prediction of low-vs high-grade prostate cancer. *Urology*, 83(6): 1362–1368.

Ankerst, DP, Straubinger, J, Selig, K, Guerrios, L, De Hoedt, A, Hernandez, J, Liss, MA, Leach, RJ, Freedland, SJ, Kattan, MW, et al., (2018). A contemporary prostate biopsy risk calculator based on multiple heterogeneous cohorts. *European Urology*, 74(2):197–203.

Assel, M, Sjoberg, DD, and Vickers, A, (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research*, 1 (1):1–7.

Austin, PC, Pencinca, M, and Steyerberg, EW, (2017). Predictive accuracy of novel risk factors and markers: a simulation study of the sensitivity of different performance measures for the Cox proportional hazards regression model. *Statistical Methods in Medical Research*, 26(3):1053–1077.

Austin, PC, (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424.

Austin, PC, (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161.

Austin, PC and Stuart, EA, (2017). Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical Methods in Medical Research*, 26 (6):2505–2525.

Austin, PC, van Klaveren, D, Vergouwe, Y, Nieboer, D, Lee, DS, and Steyerberg, EW, (2016). Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *Journal of Clinical Epidemiology*, 79(11): 76–85.

Ban, JW, Emparanza, JI, Urreta, I, and Burls, A, (2016). Design characteristics influence performance of clinical prediction rules in validation: a meta-epidemiological study. *PloS One*, 11(1):e0145779–e0145779.

Begg, CB and Greenes, RA, (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215.

Bhat, NR, Vetter, JM, Andriole, GL, Shetty, AS, Ippolito, JE, and Kim, EH, (2019). Magnetic resonance imaging-defined prostate-specific antigen density significantly improves the risk prediction for clinically significant prostate cancer on biopsy. *Urology*, 126(4):152–157.

Breslow, N, (1972). Discussion of the paper by DR Cox cited below. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2):187–220.

Breslow, NE, (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*, 43(1):45–58.

Buzoianu, M and Kadane, JB, (2008). Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach. *Statistics in Medicine*, 27(13):2453–2473.

Carbunaru, S, Nettey, OS, Gogana, P, Helenowski, IB, Jovanovic, B, Ruden, M, Hollowell, CM, Sharifi, R, Kittles, RA, Schaeffer, E, Gann, P, and Murphy, AB, (2019). A comparative effectiveness analysis of the PBCG vs. PCPT risks calculators in a multi-ethnic cohort. *BMC Urology*, 19(1):1–9.

Chen, R, Verbeek, JFM, Yang, Y, Song, Z, Sun, Y, and Roobol, MJ, (2021). Comparing the prediction of prostate biopsy outcome using the Chinese Prostate Cancer Consortium (CPCC) Risk Calculator and the Asian adapted Rotterdam European Randomized Study of Screening for Prostate Cancer (ERSPC) Risk Calculator in Chinese and European men. *World Journal of Urology*, 39(1):73–80.

Christodoulou, E, Ma, J, Collins, GS, Steyerberg, EW, Verbakel, JY, and Van Calster, B, (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110 (6):12–22.

Cook, ED, Moody-Thomas, S, Anderson, KB, Campbell, R, Hamilton, SJ, Harrington, JM, Lippman, SM, Minasian, LM, Paskett, ED, Craine, S, Arnold, KB, and Probstfield, JL, (2005). Minority recruitment to the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Clinical Trials*, 2(5):436–442.

Cook, ED, Arnold, KB, Hermos, JA, McCaskill-Stevens, W, Moody-Thomas, S, Probstfield, JL, Hamilton, SJ, Campbell, RD, Anderson, KB, and Minasian, LM, (2010). Impact of supplemental site grants to increase African American accrual for the Selenium and Vitamin E Cancer Prevention Trial. *Clinical Trials*, 7(1):90–99.

Cowley, LE, Farewell, DM, Maguire, S, and Kemp, AM, (2019). Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic and Prognostic Research*, 3(1):1–23.

Cox, DR, (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cox, DR, (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crowson, CS, Atkinson, EJ, and Therneau, TM, (2016). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706.

Cucchiara, V, Cooperberg, MR, Dall'Era, M, Lin, DW, Montorsi, F, Schalken, JA, and Evans, CP, (2018). Genomic markers in prostate cancer decision making. *European Urology*, 73(4):572–582.

Cui, T, Kovell, RC, and Terlecki, RP, (2016). Is it time to abandon the digital rectal examination? lessons from the plco cancer screening trial and peer-reviewed literature. *Current medical research and opinion*, 32(10):1663–1669.

Dahabreh, IJ, Robertson, SE, Tchetgen, EJ, Stuart, EA, and Hernán, MA, (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694.

Davis, SE, Greevy, Jr., RA, Fonnesbeck, C, Lasko, TA, Walsh, CG, and Matheny, ME, (2019). A nonparametric updating method to correct clinical prediction model drift. *Journal of the American Medical Informatics Association*, 26(12):1448–1457.

Debray, TPA, Vergouwe, Y, Koffijberg, H, Nieboer, D, Steyerberg, EW, and Moons, KGM, (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3):279–289.

Debray, TPA, Damen, JAAG, Snell, KIE, Ensor, J, Hooft, L, Reitsma, JB, Riley, RD, and Moons, KGM, (2017). A guide to systematic review and meta-analysis of prediction model performance. *British Medical Journal*, 356(1):i6460–i6460.

Drost, FH, Nieboer, D, Morgan, TM, Carroll, PR, Roobol, MJ, and the Movember Foundations Global Action Plan Prostate Cancer Active Surveillance (GAP) Consortium, (2019). Predicting biopsy outcomes during active surveillance for prostate cancer: external validation of the canary prostate active surveillance study risk calculators in five large active surveillance cohorts. *European Urology*, 76(5):693–702.

Elliot, MR, (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6):1–7.

Elliott, MR, Valliant, R, et al., (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Ferri, C, Flach, P, Hernández-Orallo, J, and Senad, A, (2005). Modifying ROC curves to incorporate predicted probabilities. In *Proceedings of the second workshop on ROC analysis in machine learning*, volume 4140, pages 33–40. International Conference on Machine Learning.

Fluss, R, Reiser, D, B.and Faraggi, and Rotnitzky, A, (2009). Estimation of the ROC curve under verification bias. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(3):475–490.

Gatz, DF and Smith, L, (1995). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. *Atmospheric Environment*, 29(11):1185–1193.

Gönen, M and Heller, G, (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970.

Guo, L, Qian, Y, and Xie, H, (2022). Assessing complier average causal effects from longitudinal trials with multiple endpoints and treatment noncompliance: An application to a study of Arthritis Health Journal. *Statistics in Medicine*. Epub ahead of print. doi:`10.1002/sim.9364`.

Haga, Y, Hato, S, Ikenaga, M, Yamamoto, K, Tsuburaya, A, Doi, K, Ikejiri, K, Hirata, T, Yamamoto, M, Ishikawa, S, and Takeuchi, H, (2018). Validation of an assessment tool: estimation of postoperative overall survival for gastric cancer. *European Journal of Surgical Oncology*, 44(4):515–523.

Harrell, FE, Califf, RM, Pryor, DB, Lee, KL, and Rosati, R, (1982). Evaluating the yield of medical tests. *Journal of America Medical Association*, 247(18):2543–2546.

Hayden, D, Pauler, DK, and Schoenfeld, D, (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics*, 61(1):305–310.

He, H and McDermott, MP, (2012). A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics*, 13(1):32–47.

He, H, Lyness, JM, and McDermott, MP, (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine*, 28(3):361–376.

Heagerty, PJ and Zheng, Y, (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Hosmer, DW, Hosmer, T, Le Cessie, S, and Lemeshow, S, (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D and Lemesbow, S, (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069.

Hosmer Jr, DW, Lemeshow, S, and Susanne, M, (2000). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Wiley, New York.

Hosmer Jr, DW, Lemeshow, S, and Sturdivant, RX, (2013). *Applied logistic regression*, volume 398. John Wiley & Sons, Hoboken, New Jersey.

Jehi, L, Ji, X, Milinovich, A, Erzurum, S, Rubin, BP, Gordon, S, Young, JB, and Kattan, MW, (2020). Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4):1364–1375.

Karim, MR and Islam, MA, (2019). *Reliability and Survival Analysis*. Springer, Singapore.

Kern, HL, Stuart, EA, Hill, J, and Green, DP, (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127.

Kiciński, M, Vangronsveld, J, and Nawrot, TS, (2011). An epidemiological reappraisal of the familial aggregation of prostate cancer: a meta-analysis. *PloS One*, 6(10):e27130–e27130.

Kim, SM, Kim, Y, Jeong, K, Jeong, H, and Kim, J, (2018). Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography*, 37(1):36–42.

Klein, EA, Thompson, IM, Lippman, SM, Goodman, PJ, Albanes, D, Taylor, PR, and Coltman, C, (2000). SELECT: the Selenium and Vitamin E Cancer Prevention Trial: rationale and design. *Prostate Cancer and Prostatic Diseases*, 3(3):145–151.

Klein, EA, Thompson, IM, Tangen, CM, Crowley, JJ, Lucia, MS, Goodman, PJ, Minasian, LM, Ford, LG, Parnes, HL, Gaziano, JM, Karp, DD, Lieber, MM, Walther, PJ, Klotz, L, Parsons, JK, Chin, JL, Darke, AK, Lippman, SM, Goodman, GE, Meyskens, FL, and Baker, LH, (2011). Vitamin E and the risk of prostate cancer: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Journal of the American Medical Association*, 306 (14):1549–1556.

Kosinski, AS and Barnhart, HX, (2003). A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present. *Statistics in Medicine*, 22(17): 2711–2721.

Kotwal, A and Schonberg, MA, (2017). Cancer screening in the elderly: a review of breast, colorectal, lung, and prostate cancer screening. *Cancer Journal*, 23(4):246–253.

Li, L, Greene, T, and Hu, B, (2018). A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Statistical Methods in Medical Research*, 27(8):2264–2278.

Lippman, SM, Goodman, PJ, Klein, EA, Parnes, HL, Thompson, IM, Kristal, AR, Santella, RM, Probstfield, JL, Moinpour, CM, Albanes, D, Taylor, PR, Minasian, LM, Hoque, A, Thomas, SM, Crowley, JJ, Gaziano, JM, Stanford, JL, Cook, ED, Fleshner, NE, Lieber, MM, Walther, PJ, Khuri, FR, Karp, DD, Schwartz, GG, Ford, LG, and Coltman, CA, (2005). Designing the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Journal of the National Cancer Institute*, 97(2):94–102.

Lippman, SM, Klein, EA, Goodman, PJ, Lucia, MS, Thompson, IM, Ford, LG, Parnes, HL, Minasian, LM, Gaziano, JM, Hartline, JA, Parsons, JK, Bearden, JD, Crawford, ED, Goodman, GE, Claudio, J, Winquist, E, Cook, ED, Karp, DD, Walther, P, Lieber, MM, Kristal, AR, Darke, AK, Arnold, KB, Ganz, PA, Santella, RM, Albanes, D, Taylor, PR, Probstfield, JL, Jagpal, TJ, Crowley, JJ, Meyskens, FL, Baker, LH, and Coltman, CA, (2009). Effect of selenium and vitamin E on risk of prostate cancer and other cancers: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Journal of the American Medical Association*, 301(1):39–51.

Liu, D and Zhou, XH, (2010). A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach. *Biometrics*, 66 (4):1119–1128.

Lobo, JM, Jiménez-Valverde, A, and Real, R, (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2): 145–151.

Louie, KS, Seigneurin, A, Cathcart, P, and Sasieni, P, (2015). Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis. *Annals of Oncology*, 26(5):848–864.

Luijken, K, Groenwold, RH, Van Calster, B, Steyerberg, EW, and van Smeden, M, (2019). Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in Medicine*, 38(18): 3444–3459.

Mehralivand, S, Shih, JH, Rais-Bahrami, S, Oto, A, Bednarova, S, Nix, JW, Thomas, JV, Gordetsky, JB, Gaur, S, Harmon, SA, et al., (2018). A magnetic resonance imaging–based prediction model for prostate biopsy risk stratification. *JAMA Oncology*, 4(5):678–685.

Metnitz, PGH, Lang, T, Vesely, H, Valentin, A, and Le Gall, JR, (2000). Ratios of observed to expected mortality are affected by differences in case mix and quality of care. *Intensive Care Medicine*, 26(7):1466–1472.

Meurer, WJ and Tolles, J, (2017). Logistic regression diagnostics: understanding how well a model predicts outcomes. *Journal of the American Medical Association*, 317(10): 1068–1069.

Miller, ME, Hui, SL, and Tierney, WM, (1991). Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8):1213–1226.

Moons, KG, Kengne, AP, Woodward, M, Royston, P, Vergouwe, Y, Altman, DG, and Grobbee, DE, (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, 98(9):683–690.

Moore, DF, (2016). *Applied survival analysis using R*. Springer, Switzerland.

Naji, L, Randhawa, H, Sohani, Z, Dennis, B, Lautenbach, D, Kavanagh, O, Bawor, M, Banfield, L, and Profetto, J, (2018). Digital rectal examination for prostate cancer screening in primary care: a systematic review and meta-analysis. *The Annals of Family Medicine*, 16(2):149–154.

National Cancer Institute, (2021). PLCO Background Information. URL `https://prevention.cancer.gov/major-programs/prostate-lung-colorectal-and-ovarian-cancer-screening-trial/background-information`.

National Cancer Institute, (2021). *PLCO Cancer Diagnosis - Prostate*. URL `https://cdas.cancer.gov/learn/plco/cancer-dx/`.

National Cancer Institute, (2021). SEER Cancer Statistics Factsheets: prostate cancer. URL `https://seer.cancer.gov/statfacts/html/prost.html`.

Newey, WK, (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2-3):201–206.

Newey, WK and McFadden, D, (1994). Chapter 36: Large sample estimation and hypothesis testing. In Engle, R and McFadden, D, editors, *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, Amsterdam, United States.

Neyman, JS, (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(translated and edited by DM Dabrowska and TP Speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10(4):1–51.

Nieboer, D, van der Ploeg, T, and Steyerberg, EW, (2016). Assessing discriminative performance at external validation of clinical prediction models. *PloS One*, 11(2):e0148820–e0148820.

Nordström, T, Akre, O, Aly, M, Grönberg, H, and Eklund, M, (2018). Prostate-specific antigen (PSA) density in the diagnostic algorithm of prostate cancer. *Prostate Cancer and Prostatic Diseases*, 21(1):57–63.

Pajouheshnia, R, Van Smeden, M, Peelen, L, and Groenwold, R, (2019). How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of Clinical Epidemiology*, 105(1):136–141.

Pearce, J and Ferrier, S, (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133(3):225–245.

Pencina, MJ and D'Agostino, RB, (2015). Evaluating discrimination of risk prediction models: the C statistic. *Journal of American Medical Association*, 314(10):1063–1064.

Pencina, MJ, D'Agostino Sr, RB, D'Agostino Jr, RB, and Vasan, RS, (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172.

Pepe, MS and Janes, H, (2013). Methods for evaluating prediction performance of biomarkers and tests. In Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, editor, *Risk Assessment and Evaluation of Predictions*, pages 107–142. Springer, New York.

Pfeiffer, RM and Gail, MH, (2017). Chapter 6: Assessment of risk model performance. In Pfeiffer, RM and Gail, MH, editors, *Absolute Risk: Methods and Applications in Clinical Management and Public Health*, pages 75–99. CRC Press, Boca Raton, Florida.

Pinsky, PF, Andriole, G, Crawford, ED, Chia, D, Kramer, BS, Grubb, R, Greenlee, R, and Gohagan, JK, (2007). Prostate-specific antigen velocity and prostate cancer gleason grade and stage. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 109(8):1689–1695.

Powers, S, McGuire, V, Bernstein, L, Canchola, AJ, and Whittemore, AS, (2019). Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population. *Statistical Methods in Medical Research*, 28(1):309–320.

Prorok, PC, Andriole, GL, Bresalier, RS, Buys, SS, Chia, D, Crawford, ED, Fogel, R, Gelmann, EP, Gilbert, F, Hasson, MA, Hayes, RB, Johnson, CC, Mandel, JS, Oberman, A, O'Brien, B, Oken, MM, Rafla, S, Reding, D, Rutt, W, Weissfeld, JL, Yokochi, L, Gohagan, JK, and PLCO Screening Trial Project Team, (2000). Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, 21(6): 273S–309S.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2013). URL `http://www.R-project.org/`.

Rahman, MS, Ambler, G, Choodari-Oskooei, B, and Omar, RZ, (2017). Review and evaluation of performance measures for survival prediction models in external validation settings. *BMC Medical Research Methodology*, 17(1):1–15.

Ramspek, CL, Jager, KJ, Dekker, FW, Zoccali, C, and van Diepen, M, (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, 14(1):49–58.

Riley, RD, Ensor, J, Snell, KIE, Debray, TPA, Altman, DG, Moons, KGM, and Collins, GS, (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *British Medical Journal*, 353(6):i3140–i3140.

Rosenbaum, PR and Rubin, DB, (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rotnitzky, A, Faraggi, D, and Schisterman, E, (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475):1276–1288.

Royston, P and Altman, DG, (2013). External validation of a Cox prognostic model: principles and methods. *BMC Medical Research Methodology*, 13(1):33–33.

Royston, P and Sauerbrei, W, (2004). A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23(5):723–748.

Rubin, DB, (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Schonlau, M, Couper, MP, et al., (2017). Options for conducting web surveys. *Statistical Science*, 32(2):279–292.

Shipe, ME, Deppen, SA, Farjah, F, and Grogan, EL, (2019). Developing prediction models for clinical use using logistic regression: an overview. *Journal of Thoracic Disease*, 11 (Suppl 4):S574–S584.

Simon, TD, Haaland, W, Hawley, K, Lambka, K, and Mangione-Smith, R, (2018). Development and validation of the Pediatric Medical Complexity Algorithm (PMCA) version 3.0. *Academic Pediatrics*, 18(5):577–580.

Snell, KIE, Archer, L, Ensor, J, Bonnett, LJ, Debray, TPA, Phillips, B, Collins, GS, and Riley, RD, (2021). External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology*, 135(7):79–89.

Song, X, Alan, S, Kellum, JA, Waitman, LR, Matheny, ME, Simpson, SQ, Hu, Y, and Liu, M, (2020). Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature Communications*, 11(1):1–12.

Southwest Oncology Group, (2021). Selenium and Vitamin E Cancer Prevention Trial (SELECT) phase III. URL `https://www.swog.org/sites/default/files/docs/2017-11/4a_S0000.pdf`.

Stevens, RJ and Poppe, KK, (2020). Validation of clinical prediction models: what does the “calibration slope” really measure? *Journal of Clinical Epidemiology*, 118(2):93–99.

Steyerberg, EW and Vergouwe, Y, (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*, 35(29): 1925–1931.

Steyerberg, EW, (2019). *Clinical prediction models*. Springer, Cham, Switzerland.

Steyerberg, EW and Harrell Jr, FE, (2016). Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology*, 69(1):245–247.

Steyerberg, EW, Harrell Jr, FE, Borsboom, GJ, Eijkemans, M, Vergouwe, Y, and Habbema, JDF, (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–781.

Steyerberg, EW, Vickers, AJ, Cook, NR, Gerds, T, Gonen, M, Obuchowski, N, Pencina, MJ, and Kattan, MW, (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*, 21(1):128–138.

Steyerberg, EW, Nieboer, D, Debray, TP, and van Houwelingen, HC, (2019). Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. *Statistics in Medicine*, 38(22):4290–4309.

Stojadinovic, M, Trifunovic, T, and Jankovic, S, (2020). Adaptation of the prostate biopsy collaborative group risk calculator in patients with PSA less than 10 ng/ml improves its performance . *International Urology and Nephrology*, 52(5):1811–1819.

Sung, H, Ferlay, J, Siegel, RL, Laversanne, M, Soerjomataram, I, Jemal, A, and Bray, F, (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3): 209–249.

SWOP, (2021). SWOP: prostate cancer risk calculator (based on ERSPC). URL `http://www.prostatecancer-riskcalculator.com/`.

Therneau, TM. *A Package for Survival Analysis in R*, (2021). URL `https://CRAN.R-project.org/package=survival`. R package version 3.2-11.

Thompson, IM, Ankerst, DP, Chi, C, Goodman, PJ, Tangen, CM, Lucia, MS, Feng, Z, Parnes, HL, and Coltman Jr, CA, (2006). Assessing prostate cancer risk: results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98(8):529–534.

Tibshirani, R, (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tolksdorf, J, Kattan, MW, Boorjian, SA, Freedland, SJ, Saba, K, Poyet, C, Guerrios, L, De Hoedt, A, Liss, MA, Leach, RJ, Hernandez, J, Vertosick, E, Vickers, AJ, and Ankerst, DP, (2019). Multi-cohort modeling strategies for scalable globally accessible prostate cancer risk tools. *BMC Medical Research Methodology*, 19(1):1–11.

Umemneku Chikere, CM, Wilson, K, Graziadio, S, Vale, L, and Allen, AJ, (2019). Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard–an update. *PLoS One*, 14(10): e0223832–e0223832.

Uno, H, Cai, T, Pencina, MJ, D'Agostino, RB, and Wei, L, (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Van Calster, B, McLernon, DJ, Van Smeden, M, Wynants, L, and Steyerberg, EW, (2019). Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):1–7.

Van Calster, B, Wynants, L, Timmerman, D, Steyerberg, EW, and Collins, GS, (2019). Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association*, 26(12):1651–1654.

van der Ploeg, T, Nieboer, D, and Steyerberg, EW, (2016). Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *Journal of Clinical Epidemiology*, 78(10):83–89.

van Klaveren, D, Gönen, M, Steyerberg, EW, and Vergouwe, Y, (2016). A new concordance measure for risk prediction models in external validation settings. *Statistics in Medicine*, 35(23):4136–4152.

van Leeuwen, PJ, Hayen, A, Thompson, JE, Moses, D, Shnier, R, Böhm, M, Abuodha, M, Haynes, AM, Ting, F, Barentsz, J, et al., (2017). A multiparametric magnetic resonance imaging-based risk model to determine the risk of significant prostate cancer prior to biopsy. *BJU International*, 120(6):774–781.

Verbeek, JFM, Nieboer, D, Steyerberg, EW, and Roobol, MJ, (2019). Assessing a patient's individual risk of biopsy-detectable prostate cancer: be aware of case mix heterogeneity and a priori likelihood. *European Urology Oncology*, 4(5):813–816.

Vergouwe, Y, Moons, KG, and Steyerberg, EW, (2010). External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology*, 172(8):971–980.

Vickers, AJ, Cronin, AM, Roobol, MJ, Hugosson, J, Jones, JS, Kattan, MW, Klein, E, Hamdy, F, Neal, D, Donovan, J, et al., (2010). The relationship between prostate-specific antigen and prostate cancer risk: the Prostate Biopsy Collaborative Group. *Clinical Cancer Research*, 16(17):4374–4381.

Wang, LY and Lee, WC, (2015). A permutation method to assess heterogeneity in external validation for risk prediction models. *PloS One*, 10(1):e0116957–e0116957.

Westreich, D, Edwards, JK, Lesko, CR, Stuart, E, and Cole, SR, (2017). Transportability of trial results using inverse odds of sampling weights. *American Journal of Epidemiology*, 186(8):1010–1014.

Yala, A, Lehman, C, Schuster, T, Portnoi, T, and Barzilay, R, (2019). A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292 (1):60–66.

Yu, W, Kim, JK, and Park, T, (2018). Estimation of Area Under the ROC Curve under nonignorable verification bias. *Statistica Sinica*, 28(4):2149–2149.

Zaytoun, O, Moussa, A, Li, J, Kattan, M, and Jones, S, (2011). Development of an improved nomogram for prediction of the outcome of the initial prostate biopsy based on readily available clinical information. *Journal of Urology*, 185(4S):e767–e767.

Zhang, Y, Alonzo, TA, and Alzheimer's Disease Neuroimaging Initiative, (2018). Estimation of the volume under the receiver-operating characteristic surface adjusting for non-ignorable verification bias. *Statistical Methods in Medical Research*, 27(3):715–739.

Zhou, XH and Castelluccio, P, (2004). Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*, 23(2):221–230.

Acknowledgements

First of all, I would like to thank my supervisor Prof. Donna Ankerst for her constant support and valuable feedback during my entire research period as her PhD student. I benefited a lot from the discussions with her. She always provided timely constructive advice on the project, which enlightened the path for the research.

Great thanks to Dr. Ruth Pfeiffer for her marvelous comments, suggestions, and nice discussions in our collaboration that helped to improve the quality of the work substantially. I would also like to thank all my colleagues in Klinikum Rechts der Isar and in the mathematical department for all the support and nice conversations. Additional thanks to the Global Challenges for Women in Math Science program.

I would also like to thank Prof. Aurélien Tellier for taking the chair of the examination committee and Prof. Byeongyeob Choi for taking time to review and assess this thesis.

At last, endless gratitude to my parents and friends for always being there and encouraging me during my whole study time. Their steady support made this PhD thesis possible.