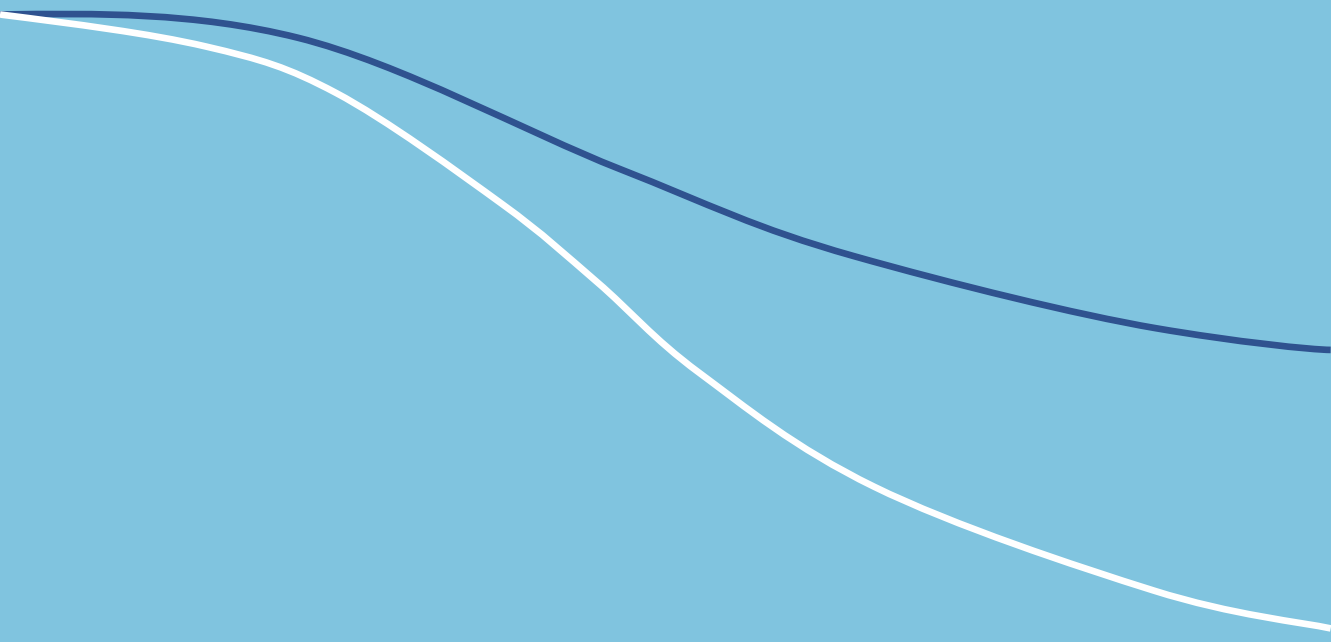
Prognostic and treatment effect modeling in medical research



Jeroen Hoogland

Prognostic and treatment effect modeling in medical research

Jeroen Hoogland

Prognostic and treatment effect modeling in medical research

PhD thesis, Utrecht University, the Netherlands

Author: Jeroen Hoogland
Cover: Jeroen Hoogland
Printed by: Gildeprint
ISBN: 978-94-6419-839-3

The research described in this thesis was financially supported by the Netherlands Organization for Health Research and Development, grant number 91215058.

Prognostic and treatment effect modeling in medical research

Predictiemodellen voor prognose en behandeleffect
binnen de medische wetenschap
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties in het openbaar
te verdedigen op

woensdag 12 juli 2023 des middags te 12.15 uur

door

Jeroen Hoogland

geboren op 1 december 1983
te Alphen a/d Rijn

Promotoren:	Prof. dr. K.G.M. Moons
	Prof. dr. M.M. Rovers
Copromotoren:	Dr. T.P.A. Debray
	Dr. J. in 't Hout
Beoordelingscommissie:	Prof. dr. S van Buuren
	Prof. dr. D.L. Oberski
	Prof. dr. K.C.B. Roes
	Dr. M. van Smeden
	Prof. dr. F.L.J. Visseren

Contents

Chapter 1

General introduction

"One of the most common and important uses for data is prediction."

Jerome H. Friedman [1]

Prediction modeling can be described as 'the process of applying a statistical or data mining algorithm to data for the purpose of predicting new or future data' [2]. Out of the vast field covered by this definition, this dissertation will focus on the use of statistical methods for prediction purposes in healthcare settings. This entails both 'regular' prediction modeling reflecting the association between a set of predictors and the outcome of interest, and causal prediction aiming to predict the effect of a modification of one or more of the predicting variables.[1]

[1] To illustrate the difference, assume that, unknown to the statistician, variability in some measure B causes variability in some measure A. Capturing the association by means of a regression of A on B will provide meaningful predictions of B for different levels of A. However, manipulating A will not change the distribution of B, and hence the model does not provide meaningful causal predictions of the effect of A.

Associative prediction modeling

Typical prediction models used in healthcare settings are of the associative kind, modeling the association between a set of predicting variables and an outcome of interest. For example, well-known clinical prediction models include the Framingham models for 10-year coronary heart disease risk [3] and general cardiovascular risk [4], the QRISK3 model predicting 10-year risk of a heart attack or stroke (`https://qrisk.org/three/`) [5], and the National Cancer Institute's Breast Cancer Risk Assessment Tool (BCRAT) for 5-year breast cancer risk predictions [6, 7]. Such risk predictions with respect to key future clinical events have clear practical relevance and play an important role in contemporary medicine. The can for instance be used to inform patients and physicians and may help to inform treatment decisions. As an example of the latter, it might be deemed desirable to restrict known effective treatment with severe side effect to those who are at a high risk of adverse outcomes based on their demographics and clinical presentation.

The road to successful implementation of such associative prediction models is well-described and researched for the associative type of prediction models [8, 9]. The general process consists of prediction model development including internal validation, external validation in independent data, and assessment of model impact in practice [10, 11]. During this process, conceptual knowledge of the underlying processes being modeled is a major factor, since it limits the amount of information that needs to be estimated from the data (*e.g.,* in terms of variables importance, interactions, and functional form) and may thereby help to diminish model uncertainty to a practically feasible level. Moreover, conceptual knowledge or study design may provide information that cannot be distilled from the data (*e.g.,* with respect to likely missing data processes, exchangeability, or possibly informative censoring). Ideally, conceptual knowledge is combined with a substantive amount of high-quality data [12, 13, 14]. As an example, this holds for all of the successfully implemented models mentioned above.

While the general process of prediction model development has been well-described, many challenges remain. One of the key challenges that arises in practical situations is the occurrence of missing data. While dealing with missing data during model development has received much attention [15, 16, 17, 18], the problem of missing data at the time of model application in practice has received only scarce attention [19]. In this dissertation, I will touch on this problem in the context of individualized risk prediction with respect to key future clinical events (**Chapter 1**). A second major challenge concerns the combination of

limited prior knowledge on the proper specification of a prediction model and limited availability of data. While novel sophisticated methods bring promising results in areas with abundant data [20], settings with relatively low numbers of observations/replications with respect to model complexity remain challenging. I will touch on this problem in the context of time-to-event data (**Chapter 2**). In addition to general prediction modeling problems, such as covariate selection for main and interaction effects and functional form, models for this type of data also have to deal with the additional challenges of censored data and possibly time-varying covariate effects. Since the amount of prior knowledge on all these modeling aspects is often limited, this is a setting in which the optimization of data-driven modeling strategies may be fruitful.

Aside from the interesting challenges that remain to be solved within associative prediction modeling, such models also have a fundamental limitation in that they do not necessarily convey the effect of interventions on the predicting variables. For instance, natural variability in systolic blood pressure is known to relate to long term risk of heart disease and therefore included in the QRISK3 model [5], but this model does not describe the effect of intervening on systolic blood pressure. To endow predictions with this desired additional meaning, the model should reflect the causal effect of systolic blood pressure on the outcome of interest, and not just their association.

Causal prediction

Causal inference aims to describes the effect of a modification of one or more variables on the outcome, and is thereby able to answer 'what if' type of questions [21, 22]. For instance, "what if I were to start a certain treatment regimen?" Within the medical domain, the preferred type of study design to answer such questions is the randomized controlled trial (RCT), which attempts to isolate the causal effect of an intervention. In such trials, patients are randomly allocated to two or more treatment regimens and followed until measurement of a clinical outcome of interest. The main goal of randomization is to ensure that the treatment groups are theoretically comparable (*i.e.*, exchangeable) with respect to factors other than treatment. Typically, these trials aim to answer questions with respect to the average effect of treatment: "Will the effect of treatment A, on average, differ from the effect of treatment B?" While well-conducted randomized controlled trials have brought about much progress, they do not directly answer the physician's most important question: "Should I administer treatment A or treatment B to *this particular individual*?" That is, the aim is to

provide individualized *causal* predictions of the outcome of interest under each of the available treatment options, with causal here referring to the idea that the differences between these predictions should be caused by differences in the chosen treatment regimen.

As will be elaborated upon in **Chapter 3**, the answer to such more individualized causal questions requires a combination of typical prediction modeling techniques and explanatory modeling techniques. The prediction modeling techniques are required to build individualized and hence potentially complicated multivariable models, while avoiding overfitting (*i.e.* still obtaining good performance in independent test data). The explanatory or causal angle is required to support the attribution of differences in predicted individualized treatment to the treatment intervention, and hence to avoid spurious association with treatment. In practice, a major challenge lies in the fact that causal effects can never be directly observed on the individual level [23]. Hence, causal prediction is a considerably more challenging task than associative prediction modeling. Nonetheless, recent years have shown an increase in the development of such methods for both randomized and observational data [24, 22]. The main focus of **Chapter 3** is on the development of prediction models for individualized treatment effect within the context of randomized trials. Subsequently, **Chapter 4** focuses on measures for their evaluation, which is still very much an open area of research. More specifically, it focuses on measures of calibration and discrimination for predicted individualized treatment effects, hence extending the typical use of calibration and discrimination measures on the level of the outcome [9, 8]. Lastly, (**Chapter 5**) describes an applied clinical study that examines possible heterogeneity in prognosis and treatment effect using data from different sources.

Outline

This dissertation describes statistical models for both typical (associative) prediction and causal prediction, with the latter focusing on individualized treatment effects in the context of randomized trial data. It touches upon modeling assumptions, model development, and model assessment procedures. Out of these, the modeling assumptions and the principles of model development are relatively well understood, but understanding of the proper evaluation procedures for causal prediction models is only just emerging.

The outline of my dissertation is as follows:

Chapter 2: 'Handling missing predictor values when validating and applying a prediction model to new patients' discusses the issue of missing data at the time of model application. If missing data are to be expected in practice, this also has implications for the handling of missing data during model validation. Most notably, care is required during internal validation.

Chapter 3: 'Regularized parametric survival modeling to improve prediction models' joins the strengths of parametric survival modeling and regularization, with a specific focus on the analysis of non-proportional hazards.

Chapter 4: 'A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint' introduces causal prediction in the context of individualized treatment effect prediction, provides guidance for the development of such models, and highlights pitfall and areas of future research.

Chapter 5: 'Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment' picks up on the performance assessment of individualized treatment effect prediction models. Model-based approaches are introduced for both discrimination and calibration purposes.

Chapter 6: 'Prognosis and prediction of antibiotic benefit in adults with clinically diagnosed acute rhinosinusitis: an individual participant data meta-analysis' is an applied study that evaluates the possible presence of treatment effect heterogeneity in a set of randomized trials. Many of the considerations encountered in earlier chapters return in this clustered data context.

Chapter 7: General discussion

Chapter 2

Handling missing predictor values when validating and applying a prediction model to new patients

Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGw, Zwinderman AH. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*, 2020; 39(25):3591-3607. DOI: 10.1002/sim.8682

Abstract
Missing data present challenges for development and real-world application of clinical prediction models. While these challenges have received considerable attention in the development setting, there is only sparse research on the handling of missing data in applied settings. The main unique feature of handling missing data in these settings is that missing data methods have to be performed for a single new individual, precluding direct application of mainstay methods used during model development. Correspondingly, we propose that it is desirable to perform model validation using missing data methods that transfer to practice in single new patients. This article compares existing and new methods to account for missing data for a new individual in the context of prediction. These methods are based on (i) submodels based on observed data only, (ii) marginalization over the missing variables, or (iii) imputation based on fully conditional specification (also known as chained equations). They were compared in an internal validation setting to highlight the use of missing data methods that transfer to practice while validating a model. As a reference, they were compared to the use of multiple imputation by chained equations in a set of test patients, because this has been used in validation studies in the past. The methods were evaluated in a simulation study where performance was measured by means of optimism corrected C-statistic and mean squared prediction error. Furthermore, they were applied in data from a large Dutch cohort of prophylactic implantable cardioverter defibrillator patients.

2.1 Introduction

An increasing number of prediction models are published in support of clinical decision making. Well-known examples in the cardiovascular domain are the QRISK3 (predicting risk of heart attack and stroke) [5] and the Seattle Heart Failure [25] models. Recently, several guidelines were published on how to perform and report prediction modeling [10, 11, 26], generally involving (i) model development, (ii) validation, and (iii) real world application. Missing data are a key issue in each of these stages. Especially the handling of missing data at the time of model development has been an active research area and multiple imputation has arisen as a general-purpose tool to account for data [27, 28]. Assuming missingness at random, multiple imputation methods allow for the use of all available data (avoiding selection bias and loss of statistical power) and at the same time account for uncertainty with respect to the missing data [27, 29, 30]. While missing data during the model development stage have attracted much attention, there is a scarcity of research on how to account for missing data during validation and real-world application of models. We propose that the methods by which missing data are handled should be an integral part of prediction model development, and be transferable to any new data, be it validation data of new individual cases.

Starting with the validation setting, prediction model validation has received considerable attention [31, 32, 33]. Its main goal is to provide empirical evidence of model performance beyond the data used for its development, ideally across different (but related) settings and populations [34]. As for prediction model development studies, validation data are usually affected by missing values. We propose that the correct way of handling missing values in validation data depends on the intended use of the to be validated model. More specifically, it depends on whether one intends to allow for missing data during model application in practice. To make the underlying rationale more clear, let's consider the use of imputation as applied independently in a set of validation data [35, 36, 37]. Use of this this strategy requires estimation of the necessary imputation models in the validation set, and thereby uses information that is not readily available in practice when a single new patient presents with missing values. That is, it uses information from other new patients (in the validation set) and in practice patients present individually. The main consequence is that the validation study approximates model performance for those with complete data. This could be in line with the intended use of the model, but the implied performance estimate is expected to be optimistic when allowing for missing

data in real life application. Also, validation performance becomes a mixture of prediction model performance and a local procedure to handle missing data. If the goal is to allow for missing data in practice, one ideally assesses prediction model performance and a transferable missing data method at the same time. Here we focus on this latter goal.

When applying previously developed prediction models in new, individual patients, accounting for missing values is not straightforward. As described above, one ideally disposes of both a prediction model and a missing data method that transfers to new individual patients. However, in practice most models do not allow for missing data, or do so by means of methods that have been shown to be problematic. Examples of prediction models enforce valid values for all predictor include implementations of the classic Framingham model (e.g. on mdcalc.com [38]) and the before mentioned Seattle Heart Failure model [25, 39]. Alternatively, some models allow for missing data on a limited set of variables and use simple imputation procedures. For example, the well-known QRISK3 model uses the average value from the development study for a measure of deprivation when geographical region is unknown (i.e. mean imputation), a conditional average based on ethnicity, age, and sex for missing values of Cholesterol/HDL ratio, blood pressure and BMI (i.e. conditional mean imputation), and zero imputation when the standard deviation of the last two blood pressure readings is missing [40]. Each of these methods has been shown to have issues in the context of model development [27], but there is no clear guidance on missing data problems in the model application stage.

As an example of the possible mismatch between model validation and model application in practice, QRISK3 validation removed all patients with unknown geographical region and used multiple imputation by chained equations to handle remaining missingness [5]. This validation does not contain any information on those with missing region and reflects performance for otherwise complete data, while the application allows for missing predictors. We have not been able to find an example in which missing data were allowed in practice and where missing data was handled consistently between validation and application.

In this paper, we propose that validation, whether internal or external, should handle missing data in a way that only depends on the development data and is applicable when making predictions for new individual patients.[1] This implies the need for missing data methods that transfer to real-life application. We

[1] As described above, when the intended use of the prediction model is to allow for missing data in practice.

consider six strategies to address missing values in individual patients when calculating a risk prediction. We compare them with the before mentioned use of (independent) multiple imputation in an internal validation setting. Our work builds on methods developed and described by Marshall et al [41] and Janssen et al [42]. We will describe their suggestions, present new methods, and describe all methods in a realistic setting including missing data in the model development data. The various methods will be illustrated with simulated data and data from an ongoing project on the prediction of mortality for primary therapy with an implantation cardioverter defibrillator (ICD) in heart failure patients at risk for cardiac arrhythmia and death (the DO-IT Registry) [43].

2.2 Methods

We consider prediction models with expectation of the form $\mathbb{E}[y_i|x_i] = g^{-1}(x_i b)$, where y_i is the outcome of patient i, x_i is the vector with values of the set of prediction variables, b is the associated vector of regression weights, and $g^{-1}(\cdot)$ is an (inverse) link-function. We here focus on the binary case, and discuss extensions to cope with censored outcomes in the applied example section.

When applying a prediction model in individual patients, several approaches can be considered to account for missing predictor values. For ease of exposition it helps to introduce some notation. First, define x_i as the partition (x_{io}, x_{im}) where x_{io} is the vector of observed predictors, and x_{im} is the vector of unobserved predictors for individual i. Analogously, define b as the partition (b_o, b_m) where b_o and b_m represent the vectors of weights of the observed and unobserved predictor variables respectively. The model of interest can then be written as $\mathbb{E}[y_i|x_{io}, x_{im}] = g^{-1}(x_{io}b_o + x_{im}b_m)$and cannot be evaluated directly due to the missing x_{im}. Several apporaches can be taken to arrive at predictions for a new individual conditional on his or her observed data only. The approaches described in the current paper can be separated into three groups based on the underlying theory. These will be shortly summarized in order to give a quick overview of the methods. To simplify notation, the subscripts will be omitted in further equations.

The first group of methods aims to find a *submodel* of the original model based on the observed covariates only. That is, the aim is to find

$$\mathbb{E}[y|x_o] = g^{-1}(x_o \check{b}_o)$$

where $\check{b}_o$ represents the vector of weights for a model conditional on the observed data only. Such a model is directly applicable for prediction purposes. The challenge for these submodel methods is to estimate $\check{b}_o$. The second group of methods integrates over the unobserved data to arrive at the predictions of interest. That is, the full model $\mathbb{E}[y|x_o, x_m])$ is integrated over the conditional distribution $g(x_m|x_o)$ as follows

$$\mathbb{E}[y|x_o] = \int \mathbb{E}[y|x_o, x_m]\, g(x_m|x_o) \partial x_m$$

where $g(x_m|x_o)$ describes the uncertainty in the unobserved data given the observed data. This *marginalization* over the unobserved data retains the original full model coefficients. The challenge for this group of methods is to estimate $g(x_m|x_o)$. The third group of methods aims to *impute* the missing covariates to enable use of the original full model, as in

$$\mathbb{E}[y|x_o, \hat{x}_m] = g^{-1}(x_o b_o + \hat{x}_m b_m)$$

where $\hat{x}_m$ contains the imputed values for the unobserved covariates. Here, the challenge lies in identification of the imputation models. All imputation methods that we considered were based on chained equations, also known as fully conditional specification [27, 16]. Imputation methods that have been shown to have issues in previous research have not been evaluated, and will not be covered in detail. These include zero imputation, mean imputation, and conditional mean imputation [27].

The methods to be described in the following sections are submodels directly estimated in the development data (method 1) and submodels based on the one-step-sweep (method 2), marginalization over the unobserved predictors (method 3) and marginalization over both the unobserved predictors and the outcome (method 4), single imputation based on chained equations (method 5) and multiple imputation based on chained equations (method 6). Each of these can be applied to new individual patients and therefore apply to both validation and application of prediction models. In addition, since it has been used in practice for validation purposes, the independent use of multiple imputation in the validation set (method 7) will be evaluated. Note however that this use of multiple imputation does not extend to new individual patients, since in that case there is not enough data to independently estimate the imputation models. Regarding terminology, development data is used to refer to the data on

which the prediction model was originally developed. Training and test data were reserved for the description of internal validation procedures to describe splitting of the development data. Importantly, note that the outcome value is always missing during model application. While it is commonly available in internal and external validation settings, the information in the observed outcomes should never be used when interest is in evaluation of model performance in real-life settings.

2.2.1 Submodel methods

The submodel approaches described by Janssen et al [42] refer to the developments of Marshall et al [41]. As described above, the underlying idea is to find the necessary submodels to cope with missing data in the application setting (*i.e.* submodels based on the observed data only). The most straightforward way to do so is to fit all necessary submodels in the development data. For a two variables example, this implies that not only the full prediction model $\mathbb{E}[y|x_1, x_2] = g^{-1}(x_1b_1 + x_2b_2)$ is fitted and reported, but also the submodels $\mathbb{E}[y|x_1] = g^{-1}(x_1\check{b}_1)$ and $\mathbb{E}[y|x_2] = g^{-1}(x_2\check{b}_2)$. The prediction for a new person with a missing x_2 value is then calculated using the $\mathbb{E}[y|x_1] = g^{-1}(x_1\check{b}_1)$ submodel. It is not difficult to estimate the submodels in the development data, but if the number of predictor variables (say, k) is large and all of them may be missing, then the number of submodels may be very large: with k predictor variables there are 2^k submodels. If $k = 15$, the number of submodels is already $32,768$ and this is not rare: both the before mentioned QRISK3 and Seattle Heart Failure model have $k \geq 15$. This direct estimation of the 2^k submodels was the first of the implemented methods.

To avoid estimation of a large number of submodels, Marshall et al [41] suggested to approximate $\check{b}$ based on the weights of the full prediction model only. Note that b may include an intercept, and hence the design matrix a corresponding unity column. The approximation starts from the assumption that the full model estimate b has a multivariate normal distribution with true mean b and covariance matrix S. Hence, by simply reporting the regression coefficients b of the full prediction model and its variance-covariance matrix S, predictions can be made for new patients, regardless of whether they are affected by missing values. Note that the estimates of b and S may also be pooled estimates over multiply imputed development data. Either way, the predictions are only based on the development data and do not require any imputation procedure for prediction for new individuals with missing data. Using the above

described partition of b as (b_o, b_m), and accordingly partitioning covariance matrix S as $\begin{pmatrix} S_{oo} & S_{om} \\ S_{mo} & S_{mm} \end{pmatrix}$, the conditional distribution of the weights of the non-missing predictor variables given the weights of the missing predictor variables is normal with approximate mean calculated with the sweeping operation as $\check{b}_o = b_o - S_{om}S_{mm}^{-1}b_m$. For instance, again using the two variable example of full model $\mathbb{E}[y|x_1, x_2] = g^{-1}(x_1b_1 + x_2b_2)$, then for a patient with missing x_2, their prediction will be based on $\mathbb{E}[y|x_1] = g^{-1}(x_1\check{b}_1$ with $\check{b}_1 = b_1 - S_{12}(1/S_{22})b_2$, where the right-hand side contains full model parameter estimates and b_1 is the estimated parameter for predictor x_1, S_{12} is the covariance between b_1 and b_2, and S_{22} is the variance of b_{22}. Interestingly, for the logistic model, predictions based on these submodels correspond one-to-one to procedures that impute x_m with the best linear predictor based on x_o, weighted by the binomial variance in the development data [41].

2.2.2 Marginalization methods: integrating over the unknown values

As described above, an alternative approach arises when we partition the vector of covariate values too, and estimate $\mathbb{E}[y|x_o]$ as follows:

$$\mathbb{E}[y|x_o] = \int \mathbb{E}[y|x_o, x_m]\, g(x_m|x_o)\partial x_m$$

All required conditional distributions can be estimated in the development data, but with large numbers of predictor variables the number of conditional distributions would again be extremely large. For this reason, we propose to estimate the joint distribution of $x = (x_o, x_m)$ in the development study, and to derive the required conditional distributions from this joint distribution. This is especially attractive when x follows the multivariate normal distribution with mean μ and covariance matrix Σ. When we partition μ as (μ_o, μ_m) and Σ accordingly as $\begin{pmatrix} \Sigma_{oo} & \Sigma_{om} \\ \Sigma_{mo} & \Sigma_{mm} \end{pmatrix}$, then the conditional distribution $g(x_m|x_o)$ has mean $\mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(x_o - \mu_o)$ and covariance $\Sigma_{mm} - \Sigma_{mo}\Sigma_{oo}^{-1}\Sigma_{om}$.

In most situations, the vector x will consist of both categorical and quantitative variables and the joint distribution will therefore almost certainly be non-normal. We hypothesize however that the normal distribution is close enough to the true joint distribution. If that is the case, then the following approach

will approximate $\mathbb{E}[y|x_o]$ to any desired degree of precision. Alternatives may involve nonparametric distributions estimated with multivariate splines [44] or copula models [45, 46].

The mean μ and covariance matrix Σ can be estimated in the development data. These $\hat{\mu}$ and $\hat{\Sigma}$ are then used for a new person i with missing data to derive the conditional distribution $\hat{g}(x_{im}|x_{io})$. We then draw a number of random vectors $\tilde{x}_{im1}, \ldots, \tilde{x}_{imj}, \ldots, \tilde{x}_{im,n_{\text{draws}}}$ from this distribution. Concatenating $\tilde{x}_{ij} = (x_{io}, \tilde{x}_{imj})$ one may calculate $\mathbb{E}[y_i|x_{io}, \tilde{x}_{imj}]$ and average over the n_{draws}:

$$\mathbb{E}[y_i|x_{io}] = \sum_{j=1}^{n_{\text{draws}}} \mathbb{E}[y|x_{io}, \tilde{x}_{imj}] \frac{\hat{g}(\tilde{x}_{imj}|x_{io})}{\sum_{r=1}^{n_{\text{draws}}} \hat{g}(\tilde{x}_{imr}|x_{io})}$$

This Monte Carlo integration approximates the integral of interest over $g(x_m|x_o)$ and was implemented as method 3 with $n_{\text{draws}} = 100$. It is based on available predictor variables and the estimated normal approximation of the joint distribution of predictors in the development data. Note that integration over is not the same as evaluation of the full prediction model at $(x_o, \mathbb{E}[g(x_m|x_o)])$.

For use of multiple imputation during model development, it has been recognized that imputation of missing x_m may also depend on y. Consequently, imputations are derived from the conditional distribution $g(x_m|x_o, y)$ [27]. If the parameters of this imputation model were known, the model could also be used to impute missing x_m given (x_o, y) in a new patient. This model is however depending on the outcome variable which is in principal not available for a new patient. One could use the entire chained-equations imputation-model from the development data and impute y too, but here we examine the possibility to integrate out y from the imputation model. This is essentially an extension of method 3 that also integrates over the outcome. In this method, we therefore use the conditional distribution $g(x_m|x_o)$ that is obtained by integrating out y:

$$g(x_m|x_o) = \int g(x_m|x_o, y) h(y|x_o) \partial y$$

If y is a binary outcome this simplifies to

$$g(x_m|x_o) = g(x_m|x_o, y = 1) h(y = 1|x_o) + g(x_m|x_o, y = 0) h(y = 0|x_o)$$

which nicely illustrates that $g(x_m|x_o)$ is obtained by averaging $g(x_m|x_o, y)$ for every possible value that y may have, but weighted with the probability that y has that particular value.

Notice that $h(y|x_o)$ is a submodel of the full prediction model, and this suggests an algorithm which is a combination of methods 2 and 3. Thus, we estimate the joint distributions $g(x|y = 0)$ and $g(x|y = 1)$ in the development data and we approximate $h(y|x_o)$ using Marshall et al's [41] suggestion (as in method 2). For a new person i with missing values of covariates in the vector x_{im}, we first sample a number of outcomes $y_{i1}, \ldots, y_{ij}, \ldots, y_{i,n_{\text{draws}}}$ from $\hat{h}(y|x_o)$ and given the sampled values $y_{ij} (j = 1, \ldots, n_{\text{draws}})$, we sample $\tilde{x}_{imj}$ from $\hat{g}(x_{im}|x_{io}, y = y_{ij})$, and $j = 1, \ldots, n_{\text{draws}}$. As with method 3 the joint distribution $g(x|y = y)$ will usually not be normal, but for the current application we approximate $g(x|y = y)$ with the multivariate normal distribution. As above, alternatives may involve nonparametric distributions estimated with multivariate splines or copula models.

2.2.3 Imputation methods

As described above, the main goal for imputations methods is to find imputations such that one can arrive at proper predictions based on the full original model. That is, the original set of regression weights (b_o, b_m) is applied to a combination of the observed and imputed values $(x_o, \hat{x}_m)$ as in

$$\mathbb{E}[y|x_o, \hat{x}_m] = g^{-1}(x_o b_o + \hat{x}_m b_m)$$

The mainstay method for multiple imputation during model development is multiple imputation by chained equations, also known as fully conditional specification [27, 29, 16]. These names refer to the typical specification where each variable has its own imputation model conditional on all the other variables (*i.e.* for the outcome given all of the x variables, for x_1 given the outcome and all other x variables, ...). That is, they are fully conditioned (on all other variables) and chained in the sense that all variables are used as both predictor and outcome. The main advantage of imputation by chained equation resides in the great flexibility that is available for the specification of each of these models, which can take any form.

It has previously been suggested that these fully conditional imputation models developed for missing data in the development dataset can also be used to

impute missing data in new patients [42]. From a methodological viewpoint, it is perfectly valid to use the previously fitted imputation model(s) in a new patient; the prediction and imputation model are considered as a unit. Although it is theoretically possible to extract the fully conditional imputation models from the development data, common software packages do not store the estimated parameters of the imputation models (*e.g.* packages like mice in R [29]; for an overview of available free and commercial statistical software for multiple imputation see Nguyen et al. [28]). To the best of our knowledge only the Amelia package in R [47], which assumes multivariate normality on the complete data, provides multiple imputation model parameters. This makes application of the imputation models to data of new patients difficult. Moreover, if the fully conditional models were available, they could not be used directly when multiple missing values are present in the new individual. This is because a fully conditional model can only be used for imputation when all predictors are known.

Two separate approaches can be taken to overcome these technical aspects. First, as proposed by Janssen et al. [42], one can simply stack the new patient below the original development data, and impute all patients together. A second possibility is to fit the required fully conditional models on the imputed development data and use these models to impute missing values in the new individual. These two methods were implemented as our method 5 and 6 respectively.

Use of the stacked imputation procedure (method 5) solves two problems. First, it does not require the imputation model parameters to be available, and second, it naturally copes with multiple missing values in the new individual. However, is also poses two new problems. First, re-running the imputation process over the combination of the entire development data and the new patient is a considerable computational burden to arrive at a single prediction. Second, a more theoretical issue is that simultaneous imputation of the development data and the new case allows sharing of information between them, while one would prefer to separate them for validation purposes. That is, the imputation model is re-estimated while it should theoretically be fixed as part of the prediction model. While this issue may only be theoretical for a single patient, the issue more clear when predictions for an entire validation set are required: the imputation models will be highly influenced by the validation data. To cope with these issues, we propose to derive the imputed development data before stacking. In this way, the imputed sets can be stored for later use (thus avoid the computational burden of the imputation process in the development data) and the imputation models are not affected by the new individual. The latter relates

to the fact that updating of the imputation models only makes use of cases with observed outcomes [27], and the new patient is thus always omitted for the necessary imputation models (*i.e.* those for which the new individual has missing values). A further issue is that imputation models used at the time of model development are based on all variables in the analysis including the outcome variable y. The outcome variable y is however missing per definition for new patients. Therefore, the chained equation approaches will automatically impute y for the new patient. This value can simply be discarded. The most important downside however, is that the original development data need to be available for every new prediction (also see Box 1 for each method's requirements). Besides computational, storage, and network issues relating to online availability of data, the most pressing issue is in limitations due to privacy regulation and data sharing limitations for many data sets.

To avoid the need for availability of the development data, we propose to derive the fully conditional model for each variable in the multiply imputed development data (method 6). This summarizes all the required information from the development data set for the future imputation process, and at the same time copes with the computational burden occurring with straightforward stacked imputation (since the imputation models are directly available and do not have to be re-estimated). Additionally, no tricks are required to avoid sharing of information between development data and new case(s). In case of missingness in the model development data, note that these fully conditional models may be pooled models over multiple imputations. Also, as for the stacked imputation, there is great flexibility in the possible classes of models to be used. For the current application, linear models were used for continuous variables and logistic regression was used for dummy coded variables. However, many more classes are conceivable and have been used successfully in multiple imputation (*e.g.* Poisson regression, multinomial regression, multi-level models) [27]. Due to estimation of the full conditional models in multiply imputed development data, the models adequately reflect the available information accounting for missing data (assuming missingness at random). Imputations for a new case can be derived iteratively in a small number of iterations. Starting from imputation of the missing x variables with the marginal means as estimated in the development data, one iterates over the full conditional models as in standard chained equation procedures. A key difference though, is that the imputation models remain fixed. First, the outcome is predicted based on the observed x variables and initial imputations for missing x variables. Second, the imputation of the first missing x variable is updated based on its fully conditional model and the

current state of the data, and so on over all other missings and repeated until convergence to the most likely imputations given the observed data (usually in < 5 iterations). Note that predicted probabilities are used in the iterative process and not the most likely binary class. Also, note that this method is essentially a simplification of traditional imputation by chained equations with the stochastic components removed. Therefore, it inherits the same theoretical limitations with respect to the relatively weak theoretical underpinnings, and assessment of its value will mainly have to come from empirical evidence [16].

2.2.4 Independent multiple imputation by chained equations for sets of patients

Lastly, while not applicable in a new patient, presence of an entire validation set allows for standard multiple imputation by chained equations as commonly used during model development. As described above, this was also the way in which the QRISK3 model was validated. A key feature of this method is that is does not allow the development data to influence validation data. However, there are at least two issues. First, the imputation method applied during validation cannot be applied in practice to new patients (hence explaining the different practical solutions implemented in for instance the QRISK3). This is only of interest when only the performance for complete cases is of interest, and the model is not to be applied in cases with missing data. Second, the imputation models are allowed to vary between development and validation set, and consequently obscure performance evaluation in the validation when transportability of the imputation procedure is of interest. Considering these issues, this method was only evaluated as a reference since it is used in practice, but it does not apply for our main goal under evaluation: application of a prediction model in a new case with missing data. If the latter is the goal of interest, we argue that it follows directly that this method should not be used for validation purposes

2.2.5 Implementation requirements

The information that is required to be able to perform these different procedures varies across the methods and ranges from just the prediction model and the variance covariance matrix of its parameters to the entire development data set. A summary of these requirements per method is available in Box 1.

Box 1: Information that needs to be available for each of the implementation missing data methods

Each of the methods to handle missing data when applying a prediction model in new patients requires additional summary statistics and or data beyond the prediction model itself. This box enlists these requirements in addition to the full model parameter vector b.

Data requirements

Method

1. *Estimation of all submodels*: requires estimated regression coefficients for all (possibly 2^k) submodels of the prediction model of interest.
2. *Submodels by means of the one-step-sweep*: only requires estimated regression coefficients and the variance-covariance matrix of developed prediction model of interest.
3. *Marginalize over missing x variables*: requires estimated means, and their variance-covariance matrix, for all variables in the development dataset that are used in the prediction model of interest.
4. *Marginalize over missing x variables and the outcome*: requirements are those for method 2 and 3 combined, where the latter are needed conditional on the outcome.
5. *Stacked multiple imputation*: requires the entire development dataset.
6. *Imputation by fixed chained equations*: requires the vector of parameter estimates for each of the fully conditional models as derived in the development dataset, as well as the means of each variable in the development data.
7. *Independent imputation by chained equations*: requires a set of test cases and can therefore not be used in case of a single new patient. This method was included for comparison in the validation setting where a set of test cases is available.

Note

In case of missing data in the development data set, multiple imputation can be used and pooled estimates can be derived for each of the required pieces of information using Rubin's rules (*e.g.* pooled model parameter estimates, variable means and variance-covariance matrices).

2.3 Simulation

2.3.1 Set-up

The set-up of the simulation study in summarized in Figure 2.1. To study the performance of the six methods we simulated data of $N = 1000$ persons with values on six predictor variables $x = (x_1, x_2, \ldots, x_6)$ and a binary outcome y. Values for x were sampled from the multivariate normal distribution with mean zero and variance 1 and a positive correlation of 0.3. Covariates x_2 and x_5 were dichotomized equal or below versus above zero, and covariates x_3 and x_6 were log-squared transformed according to $\log(0.01 + x^2)$ causing their distributions to be (left) skewed. Covariates x_1 and x_4 were not transformed. After these transformations, all continuous covariates were standardized again to have mean zero and variance 1. The binary outcome variable was modelled using a logit-link function.

Given the sampled (transformed) values for x, the probability of outcome-value $y = 1$ was calculated per person using the logit-function $\log(Odds(y = 1)) = \alpha + x\beta$, where β was chosen as $(0.8, 0.9, 1.0, 0, 0, 0)$ and α such that the relative frequency of $y = 1$ was about 30%. Given the associated probabilities $Pr(y = 1|x)$, values for y were sampled from corresponding Bernoulli distributions. For this simulation design, a (logistic) prediction model with linear additive effects of $(x_1, x_2, \ldots, x_6)$, estimated by means of maximum likelihood, leads to a c-statistic of about 0.8.

Next, we created missing data using eight scenarios. Scenarios one, two, three, and four use a completely random process with 1) 5% missing data for all variables, 2) 20% missing data for all variables, 3) 20% missing data for all variables except x_1 which had 50% missing data, and 4) 50% missing data for all variables. Scenarios five, six, seven, and eight use a missing at random process where the missingness on variable x_j depended on the observed values of y and the other observed covariates. Percentages of missing data follow the same sequence as for the missing completely at random settings. The missingness models were logistic and details are given in table 2.1.

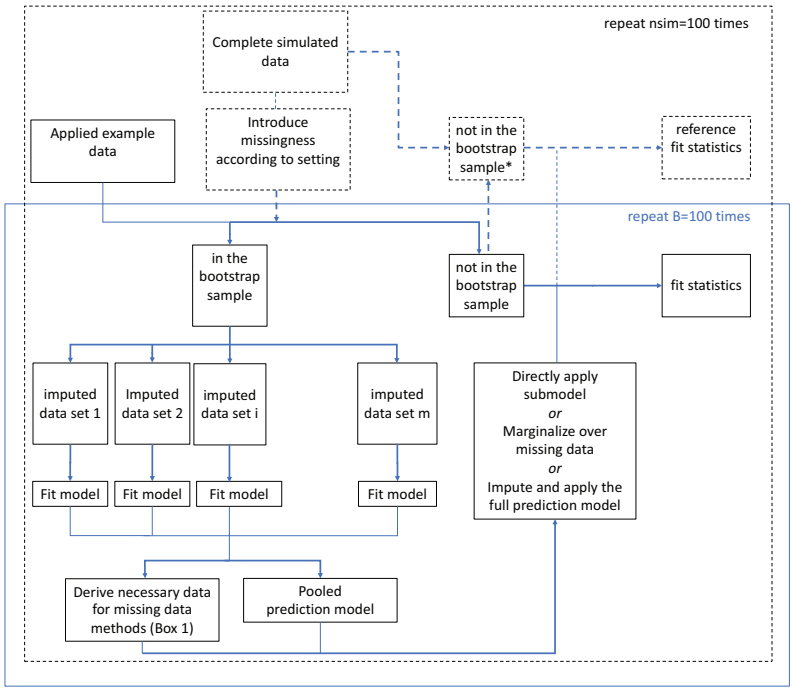


Figure 2.1: The flow of both the simulation study and applied example are shown. Parts relating only to the simulation study are shown with dashed lines. The applied example included 100 bootstrap sample evaluations. *) note that within each simulation iteration these are the same cases as the OOB samples with missing data, but with fully observed information.

Scenario	**Missing %**	**Covariates**	**Missingness mechanism**	**Missingness model**
1 MCAR	5	all x	R_{ij} ~binom(0.05)	
2 MCAR	20	all x	R_{ij} ~binom(0.20)	
3 MCAR*	20,50	x_1, x^{-1}	R_{i1} ~binom(0.50) $R_{i2}, \ldots, R_{i6}$ ~binom(0.20)	
4 MCAR	50	all x	R_{ij} ~binom(0.50)	
5 MAR	5	all x	R_{ij} ~binom(π_{ij})	$\text{logit}(\pi_{ij}) = \alpha + \beta_1 x^{-j} + \beta_2 y$ $\beta_1 = \beta_2 = 0.5$
6 MAR	20	all x	R_{ij} ~binom(π_{ij})	As for scenario (4), with α adapted for 20% missingness
7 MAR*	20,50	x_1, x^{-1}	R_{ij} ~binom(π_{ij})	$\text{logit}(\pi_{ij}) = \alpha + \beta_{1j} x^{-j} + \beta_{2j} y$ $\beta_{1,1} = \beta_{2,1} = 2.5$ $\beta_{1,2}, \ldots, \beta_{1,6} = \beta_{2,2}, \ldots, \beta_{2,6} = 0.5$
8 MAR	50	all x	R_{ij} ~binom(π_{ij})	As for scenario (4), with α adapted for 50% missingness

Table 2.1: Missingness models to create missing values in the simulated data
Abbreviation: MCAR = missing completely at random; MAR = missing at random. $R_ij = 1$ indicates that the value of covariate j in person i is missing, with $j = 1, \ldots, 6$ and $i = 1, \ldots, N$; x^{-j} is the covariate vector excluding covariate j. Parameters α for the missingness models were chosen such that the expected percentage of missing data agreed with the scenario.
*) Scenario 3 and 7 start from 2 and 6 respectively as implemented for all variables but x_1 and subsequently add the process for scenarios 4 and 8 respectively to create missing data in x_1

Given the simulated data (including introduction of missingness), a bootstrap sample was drawn with replacements and sample size equal to the full data set. Standard multiple imputation by chained equations with $m = 5$ imputed data sets was used *within* the bootstrap sample [30]. Both the pooled full (logistic) prediction model and the necessary requirements for each missing data method (see Box 1) were derived from the imputed bootstrap data. Where appropriate, these required estimates were pooled using Rubin's rules. For instance, the estimated mean and variance-covariance matrix of the variables requires for the one-step-sweep submodel method were pooled across imputations. Based on the pooled prediction model of interest and the missing data method requirements, all that needs to be estimated in the bootstrap sample is available and was applied to the out-of-bag (OOB) cases one by one. That is, predictions were derived for the OOB samples one-by-one by means of each of the missing data methods for individuals under evaluation. This one-by-one application was in line with the intended goal of the missing data methods: to provide methods that apply in practice to new individuals.

Prediction performance for these OOB cases was summarized by means of the C-statistic (as a measure of discriminative performance) and the root mean squared prediction error (rMSPE). Prediction based on multiple imputation methods were averaged. The C-statistic could be obtained directly based on the predicted values and the observed outcomes. The rMSPE was obtained based on the predicted values and the known simulated event probabilities for the OOB cases. Also, we obtained 'reference' performance measures based on complete OOB data (as illustrated in Figure 2.1). To do so, complete data was obtained for those in the OOB sample (from earlier steps in the data simulation), and the pooled prediction model was applied. This reference performance therefore corresponds to model performance in absence of missing data during model application, but already accounting for the decrease in prediction model performance caused by incomplete development data. Note that this reference is expected to be unachievable (some information is always unrecoverably lost due to missing data).

As a further comparison, independent multiple imputation in the OOB cases was evaluated (method 7). Performance measures were derived as for the methods applying to individual cases. Also, to illustrate the effect of including the outcome when performing missing data methods during model application, both stacked imputation (method 5) and independent multiple imputation (method 7) were evaluated *without* deleting the outcome in the OOB samples.

2.3.2 Simulation results

Results for discriminative performance are presented in Figure 2.2 and Supplemental Table S2.1. Mean reference performance in complete OOB samples was a C-statistic around 0.78-0.79 across missing data settings. This illustrates that standard multiple imputation by chained equations handled missing data well in the model development part of the evaluation (i.e. there was only a small decline in performance when the amount of missing data during model development increased). With respect to the missing data methods under evaluation, Figure 2.2 shows that all methods came close to reference level model performance under complete OOB data in settings with only 5% missing data. However, discrepancies began to appear when the amount of missing data increased. The one-step-sweep submodel method (method 2) was clearly less discriminative then the others. On the contrary, the approaches failing to omit the outcome information (5y and 7y) showed optimistic performance (*i.e.* higher than reference performance under complete OOB data). This clearly illustrates the need for omission of outcome information in the test set(s) of an interval validation procedures. Of the remaining methods, the 2^k submodels (method 1) and fixed chained equations (method 6) performed best and were closely followed by stacked multiple imputation (method 5). In most runs, they even performed better than independent multiple imputation in the test set (method 7). This is expected to relate to the relatively small sample size of the test data (OOB samples) with respect to the training data (bootstrap sample), which always had a ratio of approximately 1 to 1.7. Both marginalization methods (method 3 and 4) had intermediate performance.

Root mean squared prediction error results are shown in Figure 2.3. In general, performance declines as the amount of missing data increases. The comparative performance of the methods with respect to prediction error was very similar to the pattern for discriminative performance. The best performing methods are the 2^k submodel method (method 1), the fixed chained equations (method 6), and the two methods making use of the outcome information not available in practice (method 5y and 7y) that were just included for purpose of illustration.

With respect to processing times, Supplemental Figure S2.1 shows the distribution of maximum individual prediction times (including application of the missing data method) for each out-of-bag sample. As expected, stacked imputation takes longest with up to 8 seconds of processing time. However, all other methods derived predictions in less than half a second; more precisely, less than 0.3 seconds for the 2^k submodels and the marginalization approaches and less

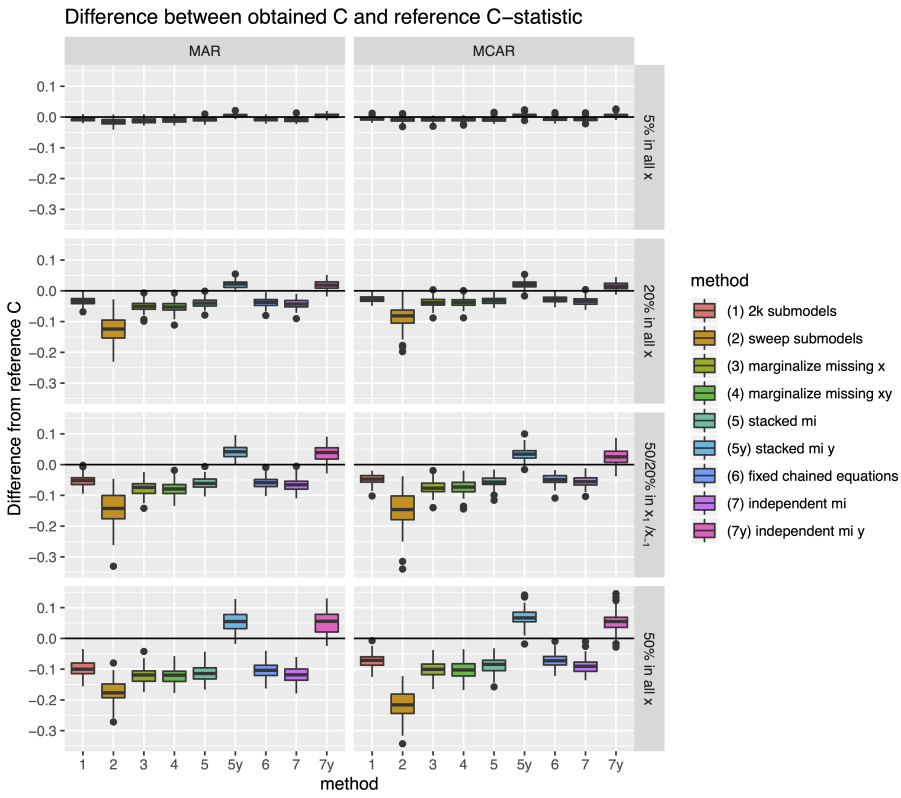


Figure 2.2: Boxplots for the difference between the estimated OOB C-statistic and reference C-statistic (as derived under complete OOB data) are shown per missing data method and missing data setting. Each simulation iteration renders an observation.

than 0.06 seconds for the one-step-sweep and fixed chained equations. These processing times illustrate applicability in practice with respect to speed of the evaluated methods and of those without stacked imputation in particular.

Beyond discriminative performance, prediction error, and processing times, Supplemental Figure S2.2 illustrates the associations between predicted probabilities derived from each of the applied methods to a those with missing data in a test

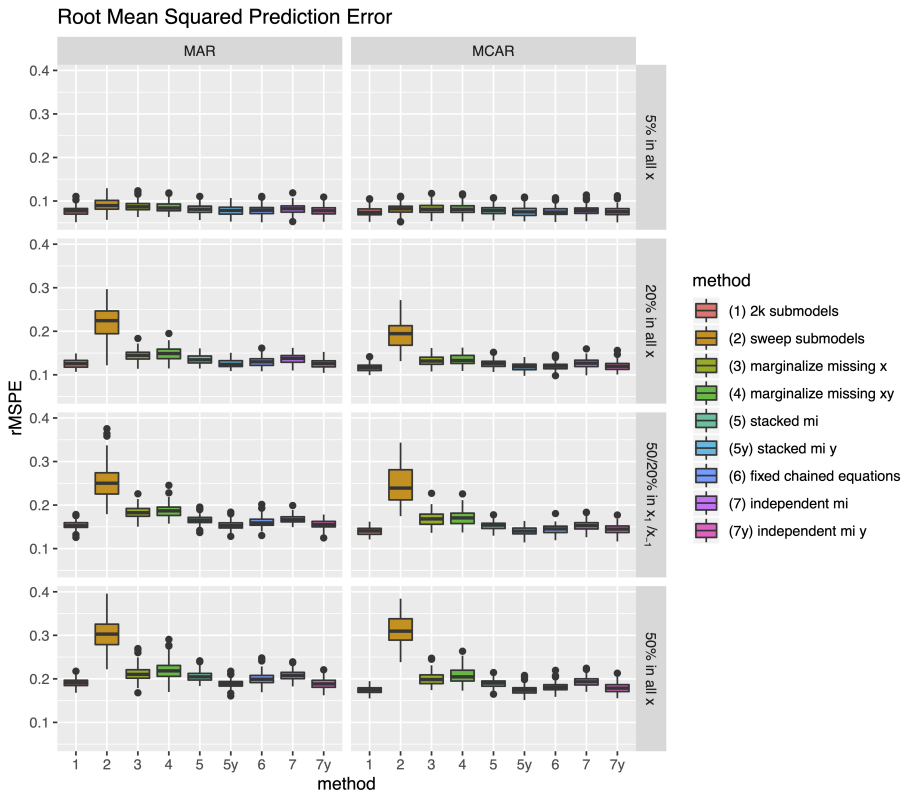


Figure 2.3: Boxplots for the average root mean squared prediction error (rMSPE) per missing data method and missing data setting. Each simulation iteration renders an observation.

set (*i.e.* OOB sample). Predicted probabilities are shown for each of the eight simulated missing data scenarios for the first simulation run. As shown, both marginalization approaches have a high correspondence across settings. The same holds for predictions based on the 2^k submodels (method 1) and those based on the fixed chained equations approach (method 6).

2.4 ICD STUDY

2.4.1 Set-up

As an empirical example, we describe the results of each of the seven methods to deal with missing data in persons in test sets with data from the DO-IT registry. In the study alongside this registry, prediction models are developed to help decision making on implantation of cardioverter defibrillators (ICD) in primary prevention patients at risk for cardiac arrhythmia and death. This registry included 1433 patients between September 2014 and June 2016 from all Dutch ICD implanting hospitals [43]. Only patients with a primary indication according to the Dutch national guidelines for ICD therapy were included. Patients were followed for occurrence of appropriate ICD therapy (defibrillator shock or antitachycardia pacing for ventricular tachyarrhythmias) or all cause death. At date of implantation, a set of 45 patient characteristics was gathered including biographic, clinical and biochemical risk factors of arrhythmia and sudden death. These included binary variables (such as sex), categorical variables (such as classes of mitral insufficiency), and continuous variables such as age, weight, NTproBNP and eGFR levels and QRS duration. Some of the continuous variables showed extremely skewed distributions.

Primary goal of the project was to develop a joint prediction model for appropriate ICD therapy and death with the total set of patient characteristics. Survival time was censored in 92% of the sample. Details are available in van Barreveld et al [43]. For the current paper, we focus only on the prediction model for all cause death. We chose to analyze these data with a Cox regression model, and therefore used a log-log link function. We used the algorithm specified in Figure 2.1 for internal validation. We performed Cox regression with AIC based backward selection of the 45 predictor variables in the imputation sets of the bootstrap training samples. Each predictor that was selected in at least half of the imputations was selected in the final model. Instead of backward selection, one could use lasso or another penalization approach to select the relevant variables; which selection algorithm is best for our data falls outside the scope of the current paper.

Inevitably, there were missing values in the set of patient characteristics. Averaged over the sample of patients and the set of characteristics, the percentage of missing values was 4.6%. However, some variables had a much higher missing data percentage, with the highest percentages for the levels of NTproBNP (60.0%) and BUN (blood urea nitrogen) (20.7%). NTproBNP also showed to

be one of the most important predictor variables.

In order to apply the methods in this survival setting with a censored outcome, several extensions were necessary for methods 4 (marginalization over x and y) and the imputations methods. These will be described here in the context of the internal validation setting of the application study. To cope with the censored outcome, we calculated martingale residuals for each person in the training sets using the Kaplan-Meier survival curve and used these residuals in the imputation models in the training sets.

For the imputation methods, the martingale residuals were included in the imputation models instead of the outcome and time-to-event. Instead of full conditional models for the event indicator and time-to-event, a linear full conditional model with the martingale residual as the outcome was used. Accordingly, the martingale residual was also used as a predictor in the full conditional models for the covariates. While improvements have been proposed [48], this was not the subject of the current study.

While these relatively simple changes suffice for the imputation methods, the extension required for method 4 is more involved. The martingale residual of person i with event or censoring at t_i has expectation zero but is usually very skewed. We nevertheless approximated the distribution of (x_i, mr_i) with the multivariate normal distribution with mean (μ_x, μ_{mr}) and partitioned covariance matrix $\begin{pmatrix} \Sigma_{xx} & \Sigma_{x,mr} \\ \Sigma_{mr,x} & \Sigma_{mr} \end{pmatrix}$ that was estimated in the training sets (averaged over the imputation sets).

Now consider persons with missing values on covariates x_m and observed values on covariates x_o. We partitioned the vector (x, mr) as (x_o, mr, x_m) with partitioned mean $(\mu_o, \mu_m r, \mu_m)$ and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{oo} & \Sigma_{o,mr} & \Sigma_{o,m} \\ \Sigma_{mr,o} & \Sigma_{mr} & \Sigma_{mr,m} \\ \Sigma_{m,o} & \Sigma_{m,mr} & \Sigma_{mm} \end{pmatrix}$$

We next approximated the distribution of (x_o, mr) negating x_m (as with method 2), with the multivariate normal distribution with mean $(\bar{\mu}_o, \bar{\mu}_{mr}) = (\mu_o, \mu_{mr}) - \Sigma_{(o,mr),m}\Sigma_{mm}^{-1}\mu_m$ and variance $\bar{\Sigma}_{o,mr|m} = \Sigma_{(o,mr)} - \Sigma_{(o,mr),m}\Sigma_{mm}^{-1}\Sigma_{m(o,mr)}$, where

$$\Sigma_{(o,mr)} = \begin{pmatrix} \Sigma_{oo} & \Sigma_{o,mr} \\ \Sigma_{mr,o} & \Sigma_{mr} \end{pmatrix}, \quad \text{and} \quad \Sigma_{(o,mr),m} = \begin{pmatrix} \Sigma_{o,m} \\ \Sigma_{mr,m} \end{pmatrix}$$

In person i with missing x_{im} values and observed values x_{io} the mean and variance of the distribution of mr_i given x_{io} was next calculated as $\bar{\bar{\mu}}_{mr} = \bar{\mu}_{mr} - \bar{\Sigma}_{mr,o|m}\bar{\Sigma}^{-1}_{o|m}(x_{io} - \bar{\mu}_o)$ and $\bar{\bar{\Sigma}}_{mr|o} = \bar{\Sigma}_{mr|m} - \bar{\Sigma}_{mr,o|m}\bar{\Sigma}^{-1}_{o|m}\bar{\Sigma}_{o,mr|m}$, where $\bar{\Sigma}_{o|m}$,$\bar{\Sigma}_{mr,o|m}$,$\bar{\Sigma}_{o,mr|m}$, and $\bar{\Sigma}_{mr|m}$ are the submatrices of $\bar{\Sigma}_{(o,mr)|m}$. We then sampled mr_i a couple of times ($n_{\text{draws}} = 100$ times) from the normal distribution with mean $\bar{\bar{\mu}}_{mr}$ and variance $\bar{\bar{\Sigma}}_{mr|o}$: $mr_{i1}, \ldots, mr_{ij}, \ldots, mr_{i,n_{\text{draws}}}$.

Given the sampled value of the martingale residual mr_{ij}, the mean and variance of the conditional distribution $(x_m|x_{io}, mr_i = mr_{ij})$ were calculated in a similar fashion as described under method 3 and we sampled then a couple of values x_m from this distribution: $x_{im1}, \ldots, x_{imj}, \ldots, x_{im,n_{\text{draws}}}$. Given the sampled values for x_m and given the observed values for x_{io} the linear predictor of the Cox regression model was calculated for patient i and averaged over the sampled values for x_m.

2.4.2 Application results

The apparent results and the internal validation results based on these survival extensions were as follows. The median number of predictor variables that was selected in the 100 bootstrap training sets was 8 (IQR 7-10). Almost all predictor variables were selected at least once, but only age, weight, mitral insufficiency category, use of diuretics, blood sodium, blood urea nitrogen, ACE inhibitor or AT-II antagonist use, and NTproBNP were selected more than 40% of the time. The average apparent c-statistic calculated in the 100 bootstrap samples 0.827 (sd 0.023) and the average c-statistics over the 100 out-of-bag samples are shown in Table 2.2. All methods showed very similar results, with the patterns of differences among methods similar to the simulations: the corrected C-statistic for the one-step-sweep submodels was relatively low and that for methods failing to ignore the outcome was relatively high. Given the relatively low proportion of missing data in the applied example, these relatively similar results across methods were expected and are in line with the simulation study results.

Method	Mean (OOB) C (sd)* in the test sets
2^k submodels	0.747 (0.034)
One-step-sweep submodel	0.736 (0.041)
Marginalization over missing x variables	0.747 (0.034)
Marginalization over missing x and y	0.747 (0.034)
Stacked multiple imputation	0.747 (0.034)
Stacked multiple imputation with y	0.764 (0.033)
Fixed chained equations	0.748 (0.033)
Independent multiple imputation	0.746 (0.034)
Independent multiple imputation with y	0.756 (0.034)

Table 2.2: Prediction performance statistics for the applied example
*mean over 100 out-of-bag samples.

2.5 Conclusion

With implementation of a prediction model there is a choice to make about whether missing values of predictor variables are accepted for a patient who wants to know their likelihood of some future outcome. If one chooses not to accept missing values in new patients we think that validation of the prediction model should be done with test sets without missing data, or using independent multiple imputation in the test data (method 7). We focused on the setting where one wants to allow for missing data in during model application in practice, and therefore in model validation as well. We propose to only use missing data methods in validation that can also be used in practice in single new patients, and have considered several ways of dealing with missing values for new patient when applying or validating a prediction model.

With respect to accuracy of predictions for new individual patients in case of missing data, use of the 2^k submodels (method 1) and use of fixed chained equations (method 6) were best in terms of corrected C-statistic and root mean squared prediction error, with only small mutual differences. Both methods abide by our two main principles: (i) the imputations should only depend on the model development data, and (ii) they should be applicable in new individual patients. Furthermore, predicted event probabilities as derived by both methods for new individuals with missing data were very highly correlated across missing data settings. However, the methods are very different in nature. The

2^k submodels method uses a different prediction model for each missing data pattern, whereas the same full prediction model is used on imputed data when applying fixed chained equations.

Of the remaining methods, marginalizing over the missing data (method 3 and 4) and use of stacked multiple imputation (method 5) showed intermediate performance with respect to the above described methods. Submodels based on the one-step-sweep (method 2) did not perform well. Importantly, our evaluation of imputation methods that fail to ignore available data on the outcome in the test set showed over-optimistic performance estimates. This also holds for use of independent multiple imputation in the test data. It is therefore key to omit outcome data in the test set when validation a model for use in practice. Interestingly, independent multiple imputation in the test set was included to show reference performance, but it was outperformed by both method 1(2^k submodels) and 6 (fixed chained equations).

Lastly, the difference between the evaluated methods were small in the applied example, which had an average percentage of missing data of 4.6%. These results were as expected when looking at the simulation study results for a relatively low proportion of missing data, and the pattern across methods was similar as well. Therefore, the difference between the different methods will only start to have a larger impact on the results when the proportion of missing data increases.

2.6 Discussion

We have evaluated two submodel methods, two marginalization methods, and two imputation methods to derive predictions for new individuals with missing data. Several of these methods show promising results, with the best performance for estimation of separate submodels based on observed covariates only (2^k submodels) and an imputation approach based on fixed chained equations. Also, computation times were extremely fast for these two methods.

A key feature of all of the evaluated approaches was that they were only based on the prediction model development data. Therefore, both the prediction model of interest and the requirements for the method to handle missing data in future individuals can be considered as a unit. We have proposed to also use these methods when validating a prediction model that is intended to cope with missing data in practice (in contrast to independent use of multiple imputation

in the validation set). To the best of our knowledge, the notion that both the prediction model and the missing data method for use in practice should be used during model validation has not been fully recognized.

Beyond these key messages, the differences among the evaluated methods are worth some discussion. Starting with the theoretical basis, both the submodel methods and marginalization methods have a firm theoretical grounding. The submodels based on observed data only are an obvious reflection of all the available information. While our implementation of the estimation of submodels leans on the missing at random assumption (due to being estimated in multiply imputed data that was imputed under that assumption), this is not strictly necessary. Mercaldo and Blume have recently implemented a pattern-mixture variant that does not need this assumption [19]. The downside is that the submodels used in their approach are more difficult and sometimes impossible to estimate. The great computational, storage, and reporting savings achieved by the one-step-sweep submodels are achieved by additional assumptions, among which the multivariate normality of prediction model coefficients. These assumptions led to a decrease in performance offsetting the benefits.

The marginalization approaches, marginalizing over the missing data, are effectively just another way to arrive at the submodel of interest by integrating out the unknown covariates. The main limiting factor for these methods is not in their theoretical basis, but in the implementation that assumed multivariate normality of the data. If the multivariate distribution of the data could be properly reflected, these methods should retain all relevant information.

The story is somewhat different for the imputation approaches which all make use of chained equations. There has long been a lack of strong theoretical grounding for the use of imputation by means of chained equations. Citing from an overview article on imputation using chained equations by White et al. [16]: "justification of the multiple imputation by chained equations procedure has rested on empirical studies rather than theoretical arguments". Nonetheless, advances have been made recently and this literature is nicely summarized in the second edition of van Buuren's monograph on missing data (Section 4.5-4.6) [17]. Here we highlight two key references. First, Hughes et al. provided conditions (compatibility and non-informative margins) on the conditional models under which chained equation based imputations are draws from the joint distribution of interest (finite-sample results) [49]. Second, Liu et al. provided asymptotic results showing that compatibility alone is sufficient as sample size tends to infinity [50]. In practice though, model compatibility is difficult to

check. In fact, citing Liu et al.[50]: "it is precisely when a joint model is difficult to obtain that iterative imputation is preferred." Regardless of the difficulty of checking these theoretical properties in practice, imputation by means of chained equations has been used effectively in many areas [17]. The main benefit of the chained equations resides in the great amount of flexibility in model specification. Basically, any model can be used, thus avoiding the possibly problematic assumption of multivariate normality. With respect to the fixed chained equations, note that they are essentially a simplified version of the standard chained equations implementations where all stochastic elements are removed: the imputation model parameters remain fixed. Also, note that it is relatively straightforward to extend the use of fixed chained equations to allow for multiple imputation. Instead of using the point estimates for the imputation model coefficients, one can sample coefficients from the estimated multivariate normal distribution of imputation model coefficients and thereby propagate the uncertainty. The main rationale for use of single imputation in the current implementation of fixed chained equations related to the interest in point predictions, which do not require propagation of uncertainty.

Beyond theoretical aspects, more practical aspects are often limiting factors in practice. These primarily relate to processing speed and data availability. For instance, use of stacked imputation, as originally proposed by Janssen et al [42], is computationally very expensive, because each new prediction requires imputation of the entire development data. Possibly even more important is that the development data has to be available at the time of prediction, which is often not possible due to privacy regulations. For instance, we are currently developing prediction models for mortality of metastatic cancers using training data of the Dutch cancer registry and test data of the Belgium cancer registry. Both data sets cannot leave their respective countries, making stacked imputation virtually impossible. All other methods can be performed based on summaries of the development data, as shown in Box 1. While these summaries can be quite extensive (such as 2^k imputation models), modern computers and mobile apps can easily store and process this amount of information.

Following the need for missing data methods applicable in practice, we have proposed that prediction model validation should also be based on these methods. The main reason for doing so is when one wants to allow for missing data in practice. If that is not the case, then use of standard multiple imputation in development and validation data separately would provide as estimate of performance when all variables are observed. Besides the intended use of the prediction model, a brief discussion of the similarity between the internal and

external validation setting is of interest. We propose that they are handled in the same way, using missing data methods that transfer to practice in the validation data (whether hold-out sample, cross-validation hold-out fold, out-of-bag samples, or truly external data). The alternative for internal validation would be to impute first and cross-validate or bootstrap later. However, in case of internal validation and use of multiple imputation, it is preferable to let the bootstrap evaluations reflect the uncertainty in estimation of the imputation models [37]. We think this argument extends to other missing data methods.

With respect to study limitations, we did not evaluate the possible use of auxiliary variables that are not included in the prediction model, but that might provide information about missing variables. If these auxiliary variables are available at the time of model developments and application, they could be envisioned to improve imputation procedures. Also, we have evaluated performance based on point predictions, but did not tough upon their uncertainty. Furthermore, since we have evaluated an internal validation setting, we have not evaluated generalizability to other settings. Just as prediction models may need updating in new populations, the required data for each of the missing data methods may also need updating for those settings. In that sense, they are just additional models and have to be treated accordingly. Lastly, the evaluated methods all assume missingness at random. When there is a strong suspicion that missing data may be missing not at random, the above described method by Mercaldo and Blume may be of interest [19].

Summarizing, the allowance for missing data when applying a prediction model to new individuals requires specific missing data methods that differ from the model development setting. We have proposed and evaluated such approaches and have shown good performance of a submodel method basing predictions on observed data only and an imputation method based on fixed chained equations. Both are feasible in practice and the choice should be made based on aspects beyond accuracy and computational burden, such as the desire for a single prediction model (as for fixed chained equations) or lack of the need for imputation (as for the submodel methods). Moreover, we have emphasized the need to use missing data methods that translate to practice during prediction model validation.

Acknowledgements

JH and JBR acknowledge financial support from the Netherlands Organisation for Health Research and Development (grant 91215058). TD acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050) and the Dutch Heart foundation (grant 2018B006). We want to thank Arthur Wilde, MD. Professor, Amsterdam UMC, for providing the DO-IT Registry data and for his comments on the manuscript. This research was supported by The Netherlands Organisation for Health Research and Development (ZonMw; grant number 91617050) and Dutch National Health Care Institute (Zorginstituut Nederland; grant number 837004009).

Data Availability

Data from the DO-IT trial [43] are not publicly available. `R` scripts to perform the simulation study, including data generation and analysis, are available for sharing.

Supplementary Material

Supplementary Figure S2.1, Figure S2.2, and Table S2.1 are included on the following pages for ease of reference.

Missing data scenario	Method	Mean training C (SD)*	Mean test (OOB) C (SD)*
no missing data		0.801 (.024)	0.793 (.021)
5% MCAR in all x		0.802 (.024)	
	Ref.		0.793 (0.021)
	m1		0.787 (0.021)
	m2		0.784 (0.021)
	m3		0.784 (0.021)
	m4		0.784 (0.021)
	m5		0.785 (0.021)
	m5y		0.798 (0.021)
	m6		0.786 (0.020)
	m7		0.785 (0.021)

	m7y		0.798 (0.022)
no missing data		0.801 (.022)	0.792 (.025)
20% MCAR in all x		0.801 (.024)	
	Ref.		0.790 (0.026)
	m1		0.763 (0.027)
	m2		0.704 (0.036)
	m3		0.753 (0.025)
	m4		0.752 (0.025)
	m5		0.758 (0.026)
	m5y		0.811 (0.027)
	m6		0.762 (0.027)
	m7		0.756 (0.027)
	m7y		0.806 (0.029)
no missing data		0.801 (.022)	0.792 (.024)
20% MCAR in all x but x_1 (50% MCAR in x_1)		0.804 (.025)	
	Ref.		0.790 (0.024)
	m1		0.743 (0.027)
	m2		0.646 (0.054)
	m3		0.717 (0.030)
	m4		0.719 (0.030)
	m5		0.734 (0.027)
	m5y		0.824 (0.029)
	m6		0.742 (0.027)
	m7		0.735 (0.027)
	m7y		0.815 (0.038)
no missing data		0.799 (.021)	0.793 (.029)
50% MCAR in all x		0.804 (.027)	
	Ref.		0.785 (0.028)
	m1		0.713 (0.027)
	m2		0.568 (0.042)
	m3		0.681 (0.032)
	m4		0.681 (0.032)
	m5		0.698 (0.027)
	m5y		0.853 (0.027)
	m6		0.712 (0.027)
	m7		0.695 (0.027)

	m7y		0.839 (0.040)
no missing data		0.802 (.024)	0.790 (.024)
5% MAR in all x		0.800 (.020)	
	Ref.		0.790 (0.025)
	m1		0.783 (0.025)
	m2		0.774 (0.025)
	m3		0.778 (0.025)
	m4		0.779 (0.025)
	m5		0.781 (0.025)
	m5y		0.794 (0.025)
	m6		0.782 (0.025)
	m7		0.781 (0.025)
	m7y		0.794 (0.026)
no missing data		0.803 (.020)	0.793 (.024)
20% MAR in all x		0.799 (.024)	
	Ref.		0.791 (0.024)
	m1		0.758 (0.025)
	m2		0.665 (0.046)
	m3		0.741 (0.024)
	m4		0.739 (0.025)
	m5		0.752 (0.024)
	m5y		0.813 (0.023)
	m6		0.754 (0.025)
	m7		0.750 (0.024)
	m7y		0.811 (0.028)
no missing data		0.805 (.020)	0.788 (.025)
20% MAR in all x but		0.803 (.028)	
x_1 (50% MAR in x_1)			
	Ref.		0.785 (0.026)
	m1		0.731 (0.027)
	m2		0.640 (0.053)
	m3		0.708 (0.031)
	m4		0.707 (0.030)
	m5		0.724 (0.028)
	m5y		0.829 (0.031)
	m6		0.725 (0.027)
	m7		0.719 (0.028)
	m7y		0.821 (0.041)

no missing data		0.804 (.021)	0.790 (.027)
50% MAR in all x		0.797 (.034)	
	Ref.		0.781 (0.029)
	m1		0.682 (0.030)
	m2		0.607 (0.034)
	m3		0.660 (0.028)
	m4		0.658 (0.029)
	m5		0.668 (0.030)
	m5y		0.837 (0.034)
	m6		0.676 (0.030)
	m7		0.664 (0.031)
	m7y		0.831 (0.049)

Table S2.1: Prediction performance statistics for the simulated data. Reference performance was derived in complete OOB data. Note that the 'no missing data' condition was evaluated prior to the introduction of each missing data scenario (hence replicated 8 times). *) mean over 100 bootstrap data-splits. Abbreviations: (Ref.) for reference performance, (m1) 2k submodels, (m2) one-step-sweep submodels, (m3) Marginalize over x_m, (m4) Marginalize over x_m and y, (m5) Stacked MI, (m5y) Stacked MI including y, (m6) Fixed chained equations, (m7) Independent MI, and (m7y) Independent MI including y.

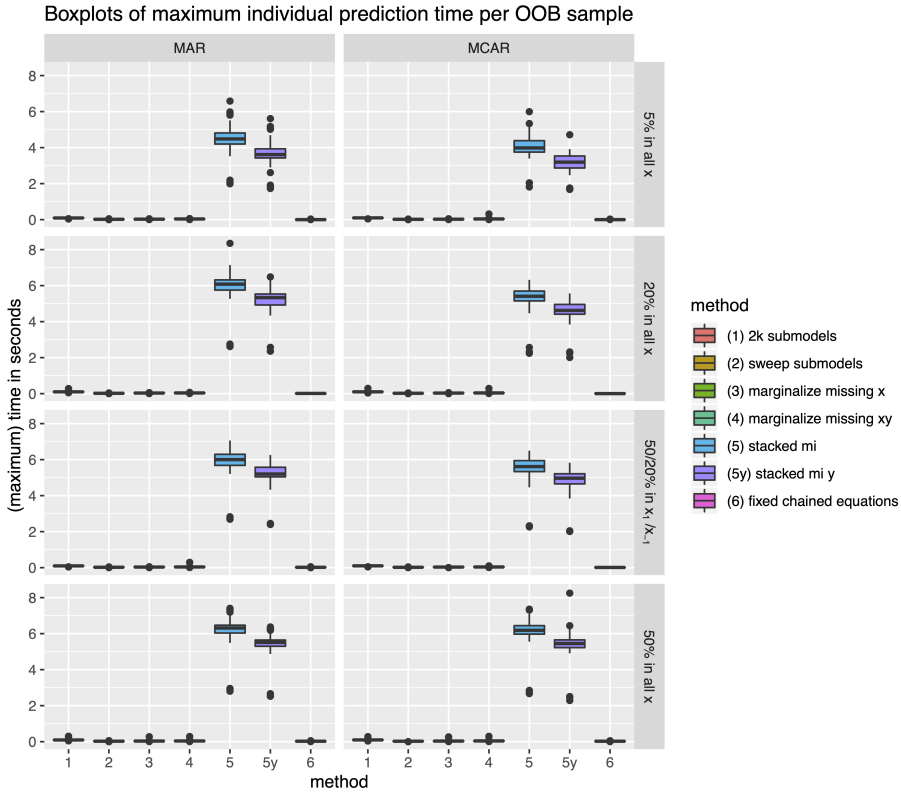
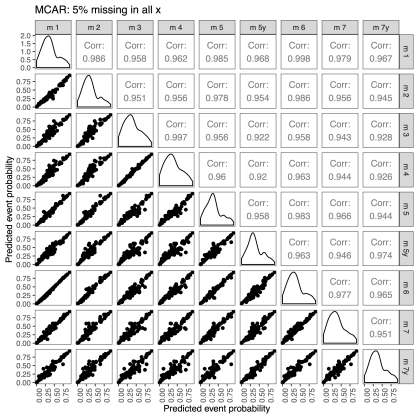
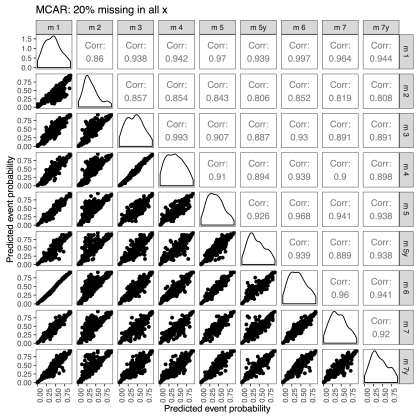


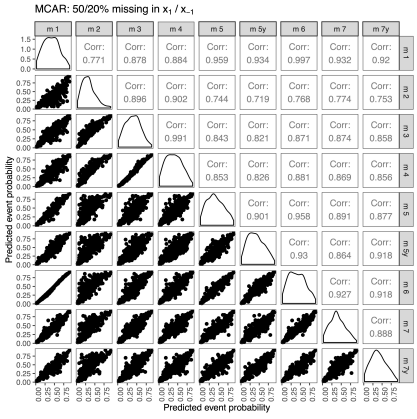
Figure S2.1: The distribution of maximum individual prediction time per OOB sample is shown per missing data method and missing data setting. Each observation is the maximum processing time to derive a prediction for an individual in an OOB sample including implementation of the missing data method.


MCAR: 5% missing in all x
m 1
m 2
m 3
m 4
m 5
m 5y
m 6
m 7
m 7y
Corr: 0.986
Corr: 0.958
Corr: 0.962
Corr: 0.985
Corr: 0.968
Corr: 0.998
Corr: 0.979
Corr: 0.967
Corr: 0.951
Corr: 0.956
Corr: 0.978
Corr: 0.954
Corr: 0.986
Corr: 0.956
Corr: 0.945
Corr: 0.997
Corr: 0.956
Corr: 0.922
Corr: 0.958
Corr: 0.943
Corr: 0.928
Corr: 0.96
Corr: 0.92
Corr: 0.963
Corr: 0.944
Corr: 0.926
Corr: 0.958
Corr: 0.983
Corr: 0.966
Corr: 0.944
Corr: 0.963
Corr: 0.946
Corr: 0.974
Corr: 0.977
Corr: 0.965
Corr: 0.951
Predicted event probability
Predicted event probability

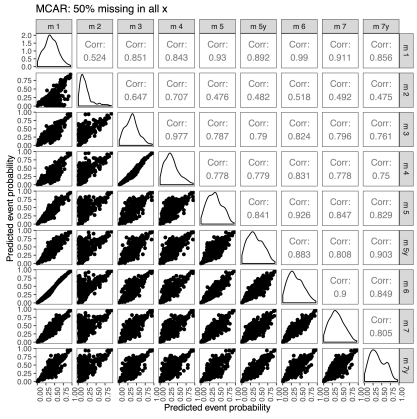
(a)


MCAR: 20% missing in all x
m 1
m 2
m 3
m 4
m 5
m 5y
m 6
m 7
m 7y
Corr: 0.86
Corr: 0.938
Corr: 0.942
Corr: 0.97
Corr: 0.939
Corr: 0.997
Corr: 0.964
Corr: 0.944
Corr: 0.857
Corr: 0.854
Corr: 0.843
Corr: 0.806
Corr: 0.852
Corr: 0.819
Corr: 0.808
Corr: 0.993
Corr: 0.907
Corr: 0.887
Corr: 0.93
Corr: 0.891
Corr: 0.891
Corr: 0.91
Corr: 0.894
Corr: 0.939
Corr: 0.9
Corr: 0.898
Corr: 0.926
Corr: 0.968
Corr: 0.941
Corr: 0.938
Corr: 0.939
Corr: 0.889
Corr: 0.938
Corr: 0.96
Corr: 0.941
Corr: 0.92
Predicted event probability
Predicted event probability

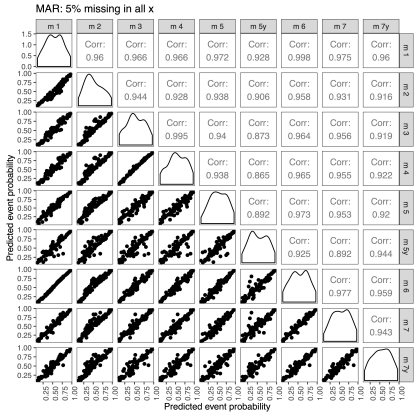
(b)


MCAR: 50/20% missing in x1 / x-1
m 1
m 2
m 3
m 4
m 5
m 5y
m 6
m 7
m 7y
Corr: 0.771
Corr: 0.878
Corr: 0.884
Corr: 0.959
Corr: 0.934
Corr: 0.997
Corr: 0.932
Corr: 0.92
Corr: 0.896
Corr: 0.902
Corr: 0.744
Corr: 0.719
Corr: 0.768
Corr: 0.774
Corr: 0.753
Corr: 0.991
Corr: 0.843
Corr: 0.821
Corr: 0.871
Corr: 0.874
Corr: 0.858
Corr: 0.853
Corr: 0.826
Corr: 0.881
Corr: 0.869
Corr: 0.856
Corr: 0.901
Corr: 0.958
Corr: 0.891
Corr: 0.877
Corr: 0.93
Corr: 0.864
Corr: 0.918
Corr: 0.927
Corr: 0.918
Corr: 0.888
Predicted event probability
Predicted event probability

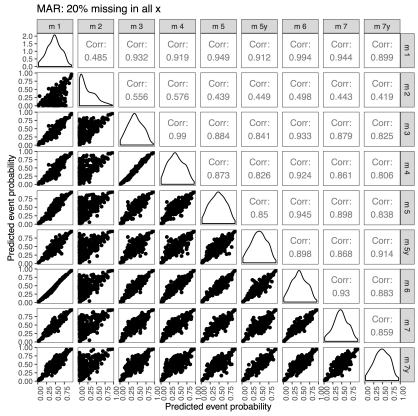
(c)


MCAR: 50% missing in all x
m 1
m 2
m 3
m 4
m 5
m 5y
m 6
m 7
m 7y
Corr: 0.524
Corr: 0.851
Corr: 0.843
Corr: 0.93
Corr: 0.892
Corr: 0.99
Corr: 0.911
Corr: 0.856
Corr: 0.647
Corr: 0.707
Corr: 0.476
Corr: 0.482
Corr: 0.518
Corr: 0.492
Corr: 0.475
Corr: 0.977
Corr: 0.787
Corr: 0.79
Corr: 0.824
Corr: 0.796
Corr: 0.761
Corr: 0.778
Corr: 0.779
Corr: 0.831
Corr: 0.778
Corr: 0.75
Corr: 0.841
Corr: 0.926
Corr: 0.847
Corr: 0.829
Corr: 0.883
Corr: 0.808
Corr: 0.903
Corr: 0.9
Corr: 0.849
Corr: 0.805
Predicted event probability
Predicted event probability

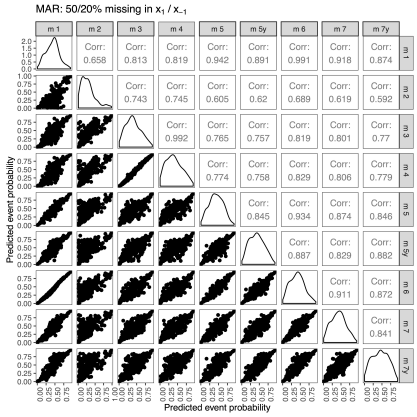
(d)



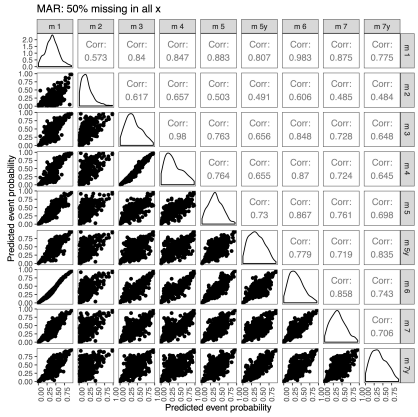
(e)



(f)



(g)



(h)

Figure S2.2: Figures a-h show the relation between the predicted event probabilities for those out of bag samples with missing data. Scatterplots for relation between predicted probabilities across the implemented methods are shown below the diagonal; their correlation is shown above the diagonal. Each sub-figure shows results for a specific missing data setting as labelled in the sub-figure titles. Predictions are shown for the first (OOB) bootstrap sample.

Chapter 3

Regularized parametric survival modeling to improve risk prediction models

Hoogland J, Debray TPA, Crowther MJ, Riley RD, IntHout J, Reitsma JB, Zwinderman AH. Regularized parametric survival modeling to improve risk prediction models. (*under revision*)

Abstract
We propose to combine the benefits of flexible parametric survival modeling and regularization to improve risk prediction modeling in the context of time to event data. Thereto, we introduce ridge, lasso, elastic net, and group lasso penalties for both log hazard and log cumulative hazard models. The log (cumulative) hazard in these models is represented by a flexible function of time that may depend on the covariates (*i.e.*, covariate effects may be time-varying). We show that the optimization problem for the proposed models can be formulated as a convex optimization problem and provide a user-friendly R implementation for model fitting and penalty parameter selection based on cross-validation. Simulation study results show the advantage of regularization in terms of increased out-of-sample prediction accuracy and improved calibration and discrimination of predicted survival probabilities, especially when sample size was relatively small with respect to model complexity. An applied example illustrates the proposed methods. In summary, our work provides both a foundation for and an easily accessible implementation of regularized parametric survival modeling and suggest that it improves out-of-sample prediction performance.

3.1 Introduction

The estimation of individualized survival probabilities is often of key interest in medical and biostatistical research [51, 8]. A suitable prediction model for this task describes survival probabilities as a function of time and the covariates of interest. In this context, fully parametric models provide a very direct and possibly parsimonious means to obtain predicted survival curves over time [52, 53]. With respect to the ubiquitous semi-parametric Cox model [54, 55], note that the way in which it elegantly avoids estimation of the baseline hazard is not a feature in this context: the baseline hazard is of key interest to obtain predicted survival probabilities. While either the Breslow estimate (of the cumulative baseline hazard) [56] or the Kalbfleisch Prentice estimate (of baseline survival) allow for survival predictions, both of these estimates involve a large number of parameters and are computationally intensive when sample size is large and/or in the presence of time-dependent effects.

A particularly flexible class of parametric survival models uses splines to model time and was introduced by Royston and Parmar [52]. This increases flexibility beyond well-known but possibly restrictive families (*e.g.* such as Weibull models), while retaining the benefits of a fully parametric model. Software implementations for Royston-Parmar model implementation are readily available (*e.g.* `stpm2` [57] in `Stata` [58] and `rstpm2` [59] in `R` [60]). Nonetheless, none of these implementations provides the means for regularization, while this has proven to be an important tool in prediction modeling to improve out-of-sample prediction accuracy [61, 20]. Key examples include ridge regression [62, 63], lasso regression [64], and elastic net regression [65]. The common idea is to put some cost on the size of model parameters (*i.e.* regression coefficients). This cost introduces a bias towards zero on the parameter level, consequently shrinking or selecting parameters, and hence reducing the variability of model predictions and thereby lowering the risk of overfitting. The objective is to find the sweet spot of this bias-variance trade-off that minimizes out-of-sample prediction error.

In this paper, we introduce regularization methods for flexible parametric survival models to aid the development of prediction models in the context of survival data. More specifically, we focus on models that are multiplicative on the hazard scale (like the well-known Cox model) or cumulative hazard scale (like the most common Royston-Parmar model). While both flexible parametric survival modeling and regularization are well-known and described in their own right, their combination is non-trivial due to the presence of constrained functions of

time described by splines (*e.g.*, the hazard and cumulative hazard function) and the allowance for interactions with these functions (*i.e.*, time-varying covariate effects). To the best knowledge of the authors, the regularization of fully parametric log hazard and log cumulative hazard modeling constitutes a novel contribution to the literature, providing the means to regularize survival models with time-varying model components that provide smooth survival estimates over time. The main aim is to increase out-of-sample accuracy of predicted survival probabilities over time in settings where sample size is limited with respect to model complexity.

In this paper, we introduce regularization methods for flexible parametric survival models to aid the development of prediction models in the context of survival data. More specifically, we focus on models that are multiplicative on the hazard scale (like the well-known Cox model) or cumulative hazard scale (like the most common Royston-Parmar model) The main aim is to increase out-of-sample accuracy of predicted survival probabilities over time in settings where sample size is limited with respect to model complexity.

The remainder of the paper is structures as follows: Sections 3.2 and 3.3 describe log cumulative hazard models and log hazard models respectively. Section 3.4 focuses on the regularization objective and details the elastic net and group lasso penalties used. Subsequently, section 3.5 details the optimization procedures with respect to the model parameters and section 3.6 discusses cross-validation for the penalty parameters. Section 3.7 illustrates the proposed methods in a simulation study and compares them to available competitive methods, and section 3.8 provides an applied example in the Veterans' Administration Lung Cancer study. Lastly, section 3.9 describes the implementation in `R` and section 3.10 provides a general discussion.

3.2 Royston-Parmar log cumulative hazard models

Royston and Parmar [52] describe a parametric model that combines proportional covariate effects with a smooth model of the log cumulative hazard as a function of log time. It has the nice property that it simplifies to the well-known Weibull proportional hazards model when the log cumulative hazard is a linear function of log time. Details are readily available in the original manuscript; here we just summarize key properties that are built upon in subsequent sections.

Let $h(t|\boldsymbol{Z})$ be the hazard at time t, conditional on $n \times p$ covariate matrix $\boldsymbol{Z}$ for n subjects and p covariates, with corresponding coefficient vector $\boldsymbol{\beta}$ expressing the log hazard ratios. The Weibull proportional hazards model (not to be confused with the accelerated failure time parametrization) can be parametrized as

$$h(t|\boldsymbol{Z}) = \zeta \nu t^{\nu-1} e^{\boldsymbol{Z\beta}} = h_0(t) e^{\boldsymbol{Z\beta}}$$

with scale parameter ζ, shape parameter ν [66], and baseline hazard $h_0(t) = \zeta \nu t^{\nu-1}$. Note that the Weibull further simplifies to the exponential distribution when $\nu = 1$, which is of importance later on. Subsequently integrating with respect to time provides the cumulative hazard formulation of the Weibull proportional hazards model.

$$\begin{aligned} H(t|\boldsymbol{Z}) &= \int_0^t \zeta \nu u^{\nu-1} e^{\boldsymbol{Z\beta}} du \\ &= \zeta t^{\nu} e^{\boldsymbol{Z\beta}} \\ &= H_0(t) e^{\boldsymbol{Z\beta}} \end{aligned}$$

Taking the (natural) logarithm on both sides gives

$$\ln H(t|\boldsymbol{Z}) = \ln\zeta + \nu \ln t + \boldsymbol{Z\beta} = \ln H_0(t) + \boldsymbol{Z\beta}$$

and shows that the Weibull log cumulative hazard form of the proportional hazards model is a linear function of log time with intercept $\ln\zeta$ and slope ν.

The Royston-Parmar proportional hazards model provides a more flexible way to model log time by means of restricted cubic splines. Starting from the log cumulative hazard form given above, log time is modeled with restricted cubic splines as

$$\ln H(t|\boldsymbol{Z}) = s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z\beta} \tag{3.1}$$

where $u = \ln(t)$ and $s(u|\boldsymbol{\alpha}, \boldsymbol{k})$ denotes the restricted cubic spline basis functions of u and their corresponding coefficients $\boldsymbol{\alpha}$ for some set of knots $\boldsymbol{k}$. The outer knots are taken to be the minimum and maximum of the observed event times, and a total of $m-2$ inner knots are set to ordered quantiles of the distribution of event times. More specifically, $s(u|\boldsymbol{\alpha}, \boldsymbol{k})$ is a linear combination of basis functions v_j, with $j \in \{1, \ldots, m\}$, and coefficients $\boldsymbol{\alpha}$ that can be written as

$$s(u|\boldsymbol{\alpha}, \boldsymbol{k}) = \alpha_0 + \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m, \tag{3.2}$$

where the definition of the basis functions depends on knots $\boldsymbol{k}$. In our implementation, we follow the definition of restricted cubic splines as provided by Harrell [67] (details are provided in the supplementary material, Part S3.1). The equivalence to the Weibull model arises when no interior knots are specified and there is only a single basis column $v_1 = u$. In presence of interior knots, the first basis function is still defined as $v_1 = u$, but further basis functions are added.

Time-dependent covariate effects One can incorporate time-dependent (or non-proportional) covariate effects by inclusion of covariate interactions with time. For instance,

$$\ln H(t|\boldsymbol{Z}) = s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z\beta} + s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa}) \tag{3.3}$$

where $s(u|\boldsymbol{\alpha}, \boldsymbol{k})$ and $\boldsymbol{Z\beta}$ are defined as in equation (3.1) and (3.2), and $s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ denotes the interaction of restricted cubic spline basis functions of u with covariate matrix $\boldsymbol{Z}_I$, where I is the subset of covariates for which a time-dependent effect is incorporated, $\boldsymbol{\kappa}$ denotes the knots for the spline of time, $\boldsymbol{\gamma}$ denotes the corresponding coefficients. For example, when two continuous covariates each interact with a restricted cubic spline representation of time with one interior knot, $s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ can be written as

$$s(u, Z_{1,2}|\boldsymbol{\gamma}, \boldsymbol{\kappa}) = \gamma_{1,1} v_1 Z_1 + \gamma_{1,2} v_2 Z_1 + \gamma_{2,1} v_1 Z_2 + \gamma_{2,2} v_2 Z_2 \tag{3.4}$$

where the γ subscripts index the covariates and spline basis functions respectively. Note that $\boldsymbol{\kappa}$ (the set of knots for interactions with time) may differ from $\boldsymbol{k}$ (the set of knots for the log cumulative baseline hazard) to allow for interactions with time that are less (or more) granular then the model for the baseline hazard. In fact, $\boldsymbol{\kappa}$ could also be a matrix $\boldsymbol{K_I}$ with different sets of knots per time-dependent covariate effect. The vector of coefficients $\boldsymbol{\gamma}$ corresponding to the example $s(u, Z_{1,2}|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ has length $m_\kappa \times q$, where m_κ is the number of basis function based on knots $\boldsymbol{\kappa}$ and q is the number of covariate columns in $\boldsymbol{Z}_I$. Further details are available in the supplementary material, Part S3.1. Also, for ease of reference, note that the parameters in equation (3.3) are grouped in log cumulative baseline hazard parameters $\boldsymbol{\alpha}$, main (proportional) effect parameters $\boldsymbol{\beta}$, and parameters relating to time-varying (non-proportional effects) $\boldsymbol{\gamma}$. Lastly, it is apparent from equation (3.3) that time-dependent effects for the log cumulative hazard model are additive on the log cumulative hazard scale (and hence multiplicative on the cumulative hazard scale). Analogously, time-dependent effects for log hazard models (next section) are additive on the log

hazard scale. That is, all time-dependent effects (and hence the $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ parameters) have an effect and interpretation that depends on the scale of the model. In contrast, the time-constant (proportional) effects with coefficients $\boldsymbol{\beta}$ have the same interpretation for both log cumulative hazard and log hazard models.

Log-likelihood The log-likelihood for both the proportional model (3.1) and the non-proportional model (3.3) is available in closed form, where the first is just a simplification of the latter. Writing $\boldsymbol{\theta} = (\boldsymbol{\alpha} : \boldsymbol{\beta} : \boldsymbol{\gamma})$, the general form of the log-likelihood is

$$l(\boldsymbol{\theta}) = \boldsymbol{\delta} \ln h(t|\boldsymbol{Z}) - H(t|\boldsymbol{Z}) \tag{3.5}$$

with $\boldsymbol{\delta}$ the vector of subject event indicators taking value 0 for (right) censored cases and 1 for events, and $H(t|\boldsymbol{Z})$ and $h(t|\boldsymbol{Z})$ given by

$$H(t|\boldsymbol{Z}) = \exp\left(s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z}\boldsymbol{\beta} + s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})\right) \tag{3.6}$$

$$h(t|\boldsymbol{Z}) = \frac{\partial}{\partial t} H(t|\boldsymbol{Z}) = H(t|\boldsymbol{Z}) \left\{s'(u|\boldsymbol{\alpha}, \boldsymbol{k}) + s'(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})\right\} \tag{3.7}$$

where $s'(\cdot)$ denotes the derivative of $s(\cdot)$ with respect to t (see the supplementary material, Part S3.1 for further detail). It is important to note that equation (3.6) should be monotone non-decreasing in time since it describes a cumulative process. Accordingly, equation (3.7) should always be non-negative. This constraint was not specifically enforced in the originally proposed optimization procedure [52], but Liu et al. later described a constrained optimization procedure to enforce non-negative hazards [59]. Nonetheless, in the most general case, a solution to the constrained optimization solution still only guarantees the constraints to hold in the development data. This is most easily seen in equation (3.7), where the contribution $s'(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$, the derivative with respect to time of the time-covariate interactions, depends on the observed data in $\boldsymbol{Z}_I$. Therefore, non-negativity of $s'(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ not only depends on the estimated parameters, but also on the observed data, and cannot be guaranteed when extrapolating beyond the development data in which the non-negative hazards constraint was enforced.

3.3 Log hazard models

In order to ensure non-negative hazards (and hence non-decreasing cumulative hazards), a straightforward alternative that automatically satisfies these constraints is to model on the log hazard scale. In that case, the linear additive part is on the log hazard scale. Since all of the components of equation (3.3) can be directly inserted here, we move directly to the general case including non-proportional effects.

$$\ln h(t|\boldsymbol{Z}) = s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z\beta} + s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa}) \tag{3.8}$$

As opposed to the log cumulative hazard models (equation (3.3)), models of the form of equation (3.8) require numerical integration to derive the cumulative hazard contributions to the log-likelihood. These cumulative hazard contributions were approximated using Gauss-Legendre quadrature [68, 53, 69].

3.3.1 Gauss-Legendre quadrature

In essence, Gauss-Legendre quadrature is just a weighted sum of q smartly chosen points at which to evaluate the function. Therefore, the likelihood contribution for individual i can be written as

$$l_i(\boldsymbol{\theta}) = \delta_i \ln h(t_i|Z_i) - \sum_{j=1}^{q} w_j h(t_{ij}|Z_i) \tag{3.9}$$

with weights w_j and evaluation time points t_{ij}. Weights for a specific j are fixed, but the time points evaluated depend on the observed time to event t_i (hence the double subscript in t_{ij}). Since the right-hand side is essentially the weighted sum of q evaluations of the hazard function at individual specific time points, an illustrative way to write the individual contributions is

$$l_i(\boldsymbol{\theta}) = \delta_i \ln h(t_i|Z_i) - \sum_{j=1}^{q} w_j e^{\boldsymbol{g}_{ij}\boldsymbol{\theta}} \tag{3.10}$$

where $\boldsymbol{g}_{ij}$ is the design vector for individual i evaluated at time point t_{ij}. That is, $\boldsymbol{g}_{ij}$ is the vector of values that corresponds to the evaluation of equation

(3.8) for individual i at time point $t = t_{ij}$. This form emphasizes that the numerical approximation of the cumulative hazard consists of a weighted sum of an exponentiated affine function $\boldsymbol{g}_{ij}\boldsymbol{\theta}$ that is linear in θ. The right choice of q for a sufficiently accurate approximation can be found in an iterative manner, increasing q until the analysis results are stable (*i.e.* specified tolerance is met).

3.4 Regularization

Regularization can be implemented by means of penalized maximum likelihood. We have implemented both an elastic net type penalty and a group lasso penalty. Section 3.6 describes tuning parameter selection based on cross-validation.

3.4.1 Elastic net

The elastic net penalty can be written as

$$P_{net}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \lambda \sum_{d=1}^{D} \omega_d \phi_d |\theta_d| + \frac{1}{2}(1 - \omega_d)\phi_d \theta_d^2 \tag{3.11}$$

with global penalty scaling parameter λ scaling the weighted sum of regression coefficient specific contributions to the penalty. The regression coefficient vector $\boldsymbol{\theta}$ has elements $d \in 1, \ldots, D$, corresponding parameter specific penalty scaling factors $\phi_d \in [0, \infty)$, and mixing factors $\omega_d \in [0, 1]$ with extremes $\omega_d = 1$ being a lasso penalty and $\omega_d = 0$ a ridge penalty. Note that in contrast to the well-known and widely applied elastic net penalty in generalized linear models ([65]), we allow for parameter specific specification of the mixing factor (as opposed to a global choice). This allows the user to combine penalties that only shrink and penalties that may also remove coefficients from the model. This is especially relevant to the survival setting. For example, it allows one to choose ridge regression for baseline (cumulative) hazard parameters (to avoid selection of individual basis functions) and a penalty that also provides parameter selection for the remaining parts of the model. With respect to the penalty scale factors ϕ_d, note that $\phi_d = 0$ equals unpenalized θ_d and that $\phi_d = \infty$ leads to $\theta_d = 0$. Setting some elements of $\boldsymbol{\phi}$ to zero could for instance be used to avoid penalization of the baseline hazard.

3.4.2 Group lasso

We implemented a group lasso penalty that can be written as

$$P_{GL}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \lambda \sum_{g=1}^{G} \omega_g \phi_g \|\boldsymbol{\theta}_g\|_2 + \frac{1}{2}(1 - \omega_g)\phi_g \|\boldsymbol{\theta}_g\|_2^2 \tag{3.12}$$

for partitions $g \in 1, \ldots, G$ of $\boldsymbol{\theta}$. Note that in this case, mixing factors ω_g and penalty scaling factors ϕ_g relate to the norms of G partitions of $\boldsymbol{\theta}$ denoted by $\boldsymbol{\theta}_g$. In addition to the usual group lasso formulation (*e.g.* [70]), and analogous to the elastic net penalty, the group lasso penalty in equation (3.12) allows for group specific ω_g, thus allowing some groups to follow a group lasso penalty and others to follow a ridge penalty. As for the elastic net case, this allows users to only shrink a subset of parameter-groups (ensuring that they stay in the model), while potentially also selecting amongst other groups of parameters (group lasso). Note that in the group lasso case, we restrict ω_g to take value in $\{0, 1\}$, but this could be extended to the entire range $[0, 1]$.

3.4.3 Survival specific nuances

Since penalization shrinks the parameters towards zero, care must be taken to specify the model such that it still makes sense under extreme penalization. In ridge, lasso, and elastic net implementations for generalized linear models, it is standard practice to avoid penalization of the intercept by centering of both outcome y and the columns of design matrix X, and to allow penalization of the remaining parameters [64, 65, 20]. However, this strategy is not directly applicable in the case of parametric survival analysis. First, centering of the outcome is not possible and the intercept therefore remains in the model and should be estimated. Therefore, our implementation treats the intercept as an unpenalized parameter. That is, the scaling factor for the intercept penalty (ϕ_1) is always equal to 0, unlike the remainder of $\boldsymbol{\phi}$. Second, a log cumulative hazard model needs at least an intercept and a slope to provide a sensible model (*i.e.* an intercept only model implies a constant log cumulative hazard). At this points, it is convenient that the first basis function of the implemented restricted cubic splines provides this slope in the form of a linear contribution of log time. Nonetheless, it may still be desirable to penalize the slope estimate. Thereto, log cumulative hazard models are estimated with a log time offset (*i.e.* slope equal to 1) whereby penalization of the slope parameter effectively shrinks towards

unity instead of zero. Consequently, the simplest model has an unpenalized intercept α_0 and an log time offset, which can be recognized as an exponential survival distribution with rate parameter e^{α_0}.

3.5 Optimization

The general optimization problem can be formulated as

$$\begin{aligned} &\text{maximize} && l_{pen}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - P(\boldsymbol{\theta}) \\ &\text{subject to} && h(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Z}) > \mathbf{0} \end{aligned} \tag{3.13}$$

where $l(\boldsymbol{\theta})$ is the appropriate form of the log-likelihood in equation (3.5) for either a log hazard or a log cumulative hazard model, $P(\boldsymbol{\theta})$ is either the elastic net penalty (equation (3.11)) or the group lasso penalty (equation (3.12)), and $h(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Z})$ denotes the hazard contributions. Note that for the latter, strict positivity could be relaxed to positivity except at event times. In order to choose the right optimization approach, we need to know more about the mathematical structure of each of the elements of the optimization problem. In this section, we show that the necessary objective functions and constraints can be written in an equivalent but convex form, such that convex optimization procedures can be used to find the global optimal value and corresponding solution(s) $\boldsymbol{\theta}^*$ [71].

3.5.1 Convexity

A particularly elegant way to establish convexity for our optimization problem is by composition of functions with well-known properties. The required composition rules for a composition $f(h_1, \ldots, h_k)$ to be convex [72], are that f is convex, and at least one of the following conditions should hold for each of the possibly vector-valued functions h_i with $i = 1, \ldots, k$

- h_i is affine,
- h_i is convex and f is increasing in argument i
- h_i is concave and f is decreasing in argument i

Alternatively, for $f(h_1, \ldots, h_k)$ to be concave, f should be concave, and at least one of the following conditions should hold for each of the possibly vector-valued functions h_i with $i = 1, \ldots, k$

- h_i is affine,

- h_i is concave and f is increasing in argument i
- h_i is convex and f is decreasing in argument i

Details on the derivation of these rules are provided elsewhere (Chapter 3 in [71]). To ease application of the rules, it is useful to write out the parts of equation (3.13) for each of the problems at hand.

Log cumulative hazard likelihood In case of a log cumulative hazard model, the likelihood contributions are already available in equation eqs. (3.5) to (3.7). For the current purpose, a more abstract formulation is more helpful. Also, the offset $\boldsymbol{u} = ln(\boldsymbol{t})$ still needs to be incorporated. Thereto, let

$$H(t|\boldsymbol{Z}) = \exp\left(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{u}\right)$$

and

$$\ln(h(t|\boldsymbol{Z})) = \boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \ln(\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha} : \boldsymbol{\beta} : \boldsymbol{\gamma})$, $\boldsymbol{X}(\boldsymbol{u})$ is the design matrix for $s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z\beta} + s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$, and $\boldsymbol{X'}(\boldsymbol{u})$ is the design matrix for $s'(u|\boldsymbol{\alpha}, \boldsymbol{k}) + s'(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$. Consequently, the log-likelihood contributions are

$$l(\boldsymbol{\theta}) = \boldsymbol{\delta}\left[\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \ln(\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta})\right] - \exp\left(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{u}\right) \tag{3.14}$$

Equation (3.14) is now composed of well-known mathematical atoms that lead to the desired result by means of the composition rules. For cases with an event, $l(\boldsymbol{\theta})$ consists of $\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta}$ which is affine, $\ln(\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta})$ which is concave (since the natural logarithm is a concave function and is taken over $\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta}$ which is affine), and $\exp\left(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{u}\right)$ which is convex (since the exponential function is convex and taken over $\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{u}$ which is affine). Hence, $l(\boldsymbol{\theta})$ is concave since it is the sum of an affine part, a concave part, and the negation of a convex (and thus also concave) part [71]. For censored cases, $l(\boldsymbol{\theta})$ simplifies to $-\exp\left(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{u}\right)$ which is concave.

Log hazard likelihood In the same line, the log-likelihood contributions for the log hazard models can be shown to be concave. Thereto, write the log-likelihood as

$$l_i(\boldsymbol{\theta}) = \boldsymbol{\delta}\boldsymbol{x}(\boldsymbol{u})_i\boldsymbol{\theta} - \sum_{j=1}^{q} w_j e^{\boldsymbol{g}_{ij}\boldsymbol{\theta}}$$

with $\boldsymbol{x}(\boldsymbol{u})_i$ the design matrix vector for individual i for $s(u|\boldsymbol{\alpha}, \boldsymbol{k}) + \boldsymbol{Z\beta} + s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$. Then the individual contributions $l_i(\boldsymbol{\theta})$ are concave since $\boldsymbol{x}(\boldsymbol{u})_i\boldsymbol{\theta}$

is affine and $-\sum_{j=1}^{q} w_j e^{\boldsymbol{g}_{ij}\boldsymbol{\theta}}$ is concave. The latter follows from the fact that $\sum_{j=1}^{q} w_j e^{\boldsymbol{g}_{ij}\boldsymbol{\theta}}$ is the sum of weighted exponentials of the affine parts $\boldsymbol{g}_{ij}\boldsymbol{\theta}$ and thus the sum of convex functions and hence itself also convex [71].

Penalty functions Both the elastic net penalty and the group lasso penalty are weighted sums of norms and/or quadratic parts and therefore weighted sums of convex functions and hence also convex.

Constraints this just leaves the constraints $h(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Z})$ which are trivial in the log hazard case and can be further simplified in the log cumulative hazard case. In the latter case, $h(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Z}) = \exp(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta})(\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta}) > \mathbf{0}$. Since $\exp(\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\theta}) > \mathbf{0}$ is always satisfied, the constraint can be simplified to $\mathbf{1} + \boldsymbol{X'}(\boldsymbol{u})\boldsymbol{\theta} > \mathbf{0}$ which is affine.

Conclusion Since the log-likelihood contributions are both concave in θ, their negation is convex; also, both penalty functions are convex, and the necessary constraints are affine. Therefore, the optimization problem can always be formulated as a convex optimization problem:

$$\begin{aligned} &\text{minimize} && -l_{pen}(\boldsymbol{\theta}) = -l(\boldsymbol{\theta}) + P(\boldsymbol{\theta}) \\ &\text{subject to} && -h(\boldsymbol{u}|\boldsymbol{\theta}, \boldsymbol{Z}) \leq \mathbf{0} \end{aligned} \tag{3.15}$$

Consequently, the proposed regularized parametric survival models have a global optimal value for fixed values of $\boldsymbol{\omega}$ and $\boldsymbol{\phi}$ which is obtained for solution $\boldsymbol{\theta}^*$.

3.5.2 Solver

The optimization problem can be formulated as a convex optimization problem for all of the proposed regularized parametric survival models. Therefore, efficient software is available for the optimization [73, 71] and is easily accessible by means of `R` package `CVXR` [72]. More specifically, `CVXR` provides a user-friendly interface that transforms the standard convex programming form of the problem into a second-order cone program, that can subsequently be solved with interior-point solver ECOS (embedded conic solver) [73].

3.6 Cross-validation

The choice of tuning parameters $\boldsymbol{\omega}$ and λ can be informed by a grid search using cross validation, repeated cross validation, or bootstrapping. The log-likelihood or deviance can be used as a measure of out-of-sample performance

[70]. However, log cumulative hazard models do not necessarily provide valid out-of-sample log-likelihoods since the non-negative hazards constraint can only be enforced within the development data. Therefore, in the context of log cumulative hazard models, we chose to optimize the objective function $l_{pen}(\boldsymbol{\theta})$ in the selected cases while enforcing the non-negative hazards constraint in the whole sample. This enforced non-negative hazards for out-of-sample predictions and hence valid out-of-sample log-likelihood contributions in the context of cross-validation (or other resampling algorithms).

3.7 Simulation study

3.7.1 Data generating mechanism

To simulate survival data, we followed a proposal by Crowther and Lambert and simulate from a two-component parametric mixture [74]. The main motivation was to generate survival data that are sufficiently complex to resemble real data, and at the same time avoid that any of the models under evaluation contains the exact data generating mechanism. Specifically, we sampled from a two-component mixture Weibull distribution that was additive on the survival scale. Clearly described details on the derivation are available elsewhere [74], so we only re-state the general form of the baseline hazard function

$$h_0(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1} p_{mix} e^{-\lambda_1 t^{\gamma_1}} + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p_{mix}) e^{-\lambda_2 t^{\gamma_2}}}{p_{mix} e^{-\lambda_1 t^{\gamma_1}} + (1-p_{mix}) e^{-\lambda_2 t^{\gamma_2}}} \tag{3.16}$$

This baseline hazard can be combined with time-independent (proportional) or time-dependent covariate effects. For our data generating mechanism, the Weibull mixture parameters were set to $\lambda_1 = 0.21, \lambda_2 = 0.05, \gamma_1 = 1.1, \gamma_2 = 1.4$ and $p_{mix} = 0.4$ and cases were censored administratively at time = 30. This mixture describes a non-monotone hazard function that first increases and subsequently decreases before stabilizing. Eleven covariates were simulated from a multivariate standard normal distribution with pair-wise correlations set to 0.25. The true main effect coefficients for these eleven covariates were $0, 0, 0.5, -0.5, 0.25, -0.25, 0.125, -0.125, 0.625, -0.625$ and 0.5 respectively. The effects of the first three covariates varied with time according to 0.9^t with coefficients $-1, 0.75$ and -0.5. Combining the time-constant and time-varying effects, the log hazard ratio of the first and second covariates started at -1 and 0.75

respectively and diminished over time, and the log hazard ratio for the third covariate started at 0 and its effect increased over time to 0.5. The supplementary material (Part S3.2) visualizes the baseline hazard and time varying effects corresponding to the data generating mechanism. A total of 110,000 observations were sampled based on this mechanism. A fixed set of 10,000 was used for external validation purposes. Development samples were sampled from the remaining 100,000 observations.

3.7.2 Simulation settings

A total of 500 simulation runs was performed for four development sample size settings: 100, 250, 500, and 1000. In each simulation run, all survival models were fitted on the development sample and evaluated in the independent validation sample. For all modeling purposes, the 11^{th} covariate was considered to be unmeasured to provide more realistic scenarios, and thus not included in the models.

3.7.3 Survival modeling methods

Ten different modeling techniques were compared.

1. Regularized log hazard model including time-varying effects (RegHazTV): regularized log hazard models with the log baseline modeled with a restricted cubic spline with 5 degrees of freedom, 10 linear main effects (*i.e.* for each measured covariate), and including interactions with log time by means of a 2 degrees of freedom restricted cubic spline for all ten covariates. The log baseline hazard and main effect parameters were penalized with a ridge penalty, and the time-varying effects with a group lasso penalty with separate groups for each covariate. With respect to the time-varying effects (*.i.e.* interactions with spline basis functions), the group lasso penalty ensures that coefficients belonging to the same spline transformation are simultaneously zero or non-zero.

2. Regularized log cumulative hazard model including time-varying effects (RegCumHazTV): same as (1), but on the log cumulative hazard scale. NB: the interactions with time are therefore also on the log cumulative hazard scale and hence differ from the specification in (1).

3. Cox proportional hazards model (CoxPH): a Cox proportional hazards model with 10 linear main effects. Predicted survival was derived based on

the Cox model and the corresponding Breslow estimate of the cumulative baseline hazard.

4. Cox model including time-varying effects (CoxTV): same as (3), but allowing for time-varying effects as a function of a 2 degrees of freedom restricted cubic spline of log time. To encode these time-varying effects, the data set was transformed into start-stop format with splits at all percentiles of the observed event times and subsequent derivation of the covariate-time interaction columns [75].

5. Cox proportional hazards lasso model (CoxPHlasso): same as (3), but with a lasso penalty on all parameters.

6. Cox time-varying effects ridge model (CoxTVridge): same as (4), but with a ridge penalty on all parameters. Note that a regular lasso penalty is not directly applicable due to the presence of spline components.

7. Royston-Parmar proportional hazards model (RPrcsPH): a proportional Royston-Parmar model (*i.e.* log cumulative hazard model) as implemented by Liu et al. [59], with a 5 degrees of freedom natural cubic spline for the log cumulative baseline hazard and 10 linear main covariate effects.

8. Royston-Parmar time-varying effects model (RPrcsTV): same as (7), but including interactions with log time by means of a 2 degrees of freedom restricted cubic spline for all ten covariates.

9. Royston-Parmar proportional hazards model (RPssPH): a proportional Royston-Parmar model (*i.e.* log cumulative hazard model) as implemented by Liu et al. [59], with a smoothing spline for the log cumulative baseline hazard and 10 linear main covariate effects.

10. Royston-Parmar time-varying effects model (RPssTV): same as (9), but including interactions with log time by means of a smoothing spline for all ten covariates.

Certain groups of methods can be distinguished within these 10 methods. For instance, we will refer to methods 1, 2, 4, 6, 8, and 10 as methods that allow for time-varying effects (TV), and to the complementary set of methods 3, 5, 7, and 9 as proportional hazards methods (PH). In addition, the methods can be grouped into methods that incorporate regularization on the size of the model parameters (methods 1, 2, 5, and 6) and methods that do not (methods 3, 4,

7-10).[1] For all regularized models, the optimal value of the penalty parameter λ was estimated by means of 10-fold cross-validation minimizing the deviance.

3.7.4 Performance measures

Prediction performance with respect to predicted survival probabilities was evaluated by means of root mean squared prediction error (rMSPE) as evaluated at the observed time-to-event for all individuals in the external validation data. Since this is a simulation setting, the true survival probabilities were used as the reference. Additionally, rMSPE was evaluated at the fixed time points $2.5, 5, 7.5, 10, 20$ and 30. Likewise, both time-average and fixed time point discriminative performance was evaluated against the true survival probabilities by means of the C-statistic [8]. Based on the observed outcomes in the validation data, calibration was assessed at the fixed time points using graphical calibration curves and integrated calibration indices based on hazard regression [76].

3.7.5 Simulation study results

Figure 3.1 shows that the time-averaged rMSPE of the proposed regularized models (RegHazTV and RegCumHazTV) were amongst the best performing methods in all sample size settings. In the smallest sample size setting (N=100), prediction accuracy of RegHazTV and RegCumHazTV slightly outperformed modeling methods assuming proportional hazards, and clearly outperformed other time-varying effects methods. With increasing sample size (N=250), the other time-varying effects methods start to catch up with the proportional hazards models. Further increase in sample size (N=500) shows that the possibility of time-varying effects models to more fully capture the data generating mechanism generally overcomes their tendency to overfit: all of the time-varying effects models outperform the proportional hazards methods. The final increase in sample size up to N=1000 shows that the non-regularized time-dependent effects models (CoxTV, RPrcsTV, RPssTV) start to catch up with the proposed regularized models.

Figure 3.2 shows rMSPE results over time. In line with the time-average results, the proposed regularized parametric methods performed well across all sample

[1] Note that methods 9 and 10 implement penalization on the second derivative of smooth functions over time (*i.e.* log cumulative baseline hazard and time-varying effect), which differs from what we refer to as regularization. More specifically, penalizing non-smoothness simplifies functional form towards linearity, while regularization reduces coefficients towards zero.

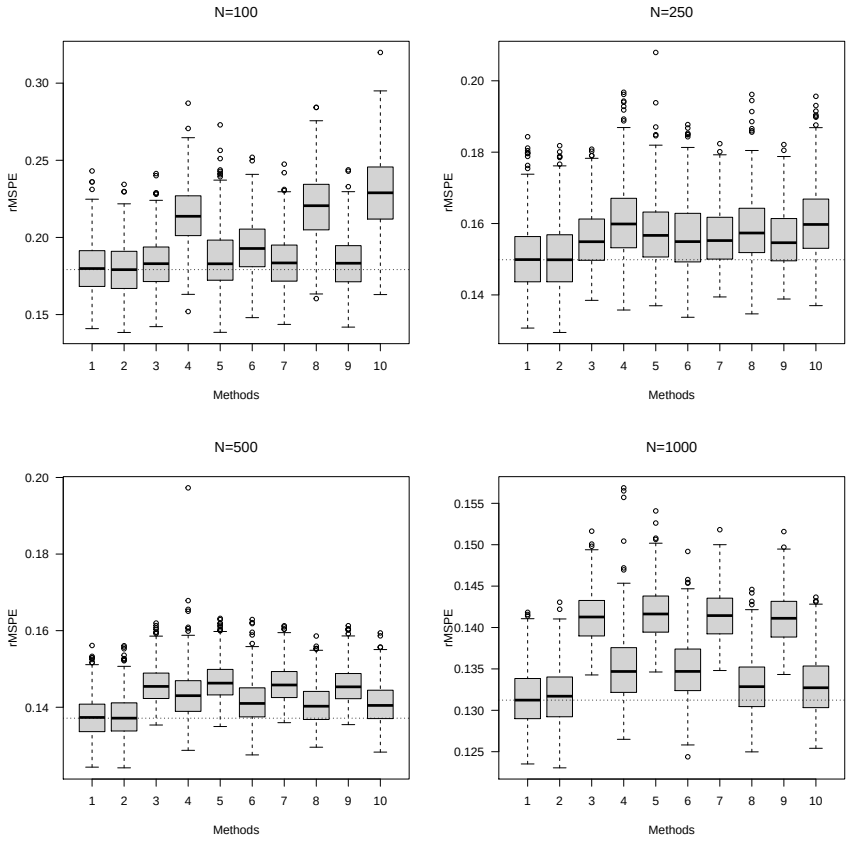


Figure 3.1: Boxplots of root mean squared prediction error for each of the methods as enumerated in section 3.7.3: (1) RegHazTV, (2) RegCumHazTV, (3) CoxPH, (4) CoxTV, (5) CoxPHlasso, (6) CoxTVridge, (7) RPrcsPH, (8) RPrcsTV, (9) RPssPH, (10) RPssTV. Boxes cover the interquartile range and have a solid bar showing the median; whiskers extend to 1.5 times the interquartile range. The horizontal dotted line crosses the median rMSPE for the best performing method in a specific sample size setting. NB: The scale of the y-axis differs between subfigures to enhance visual clarity of the differences within scenarios.

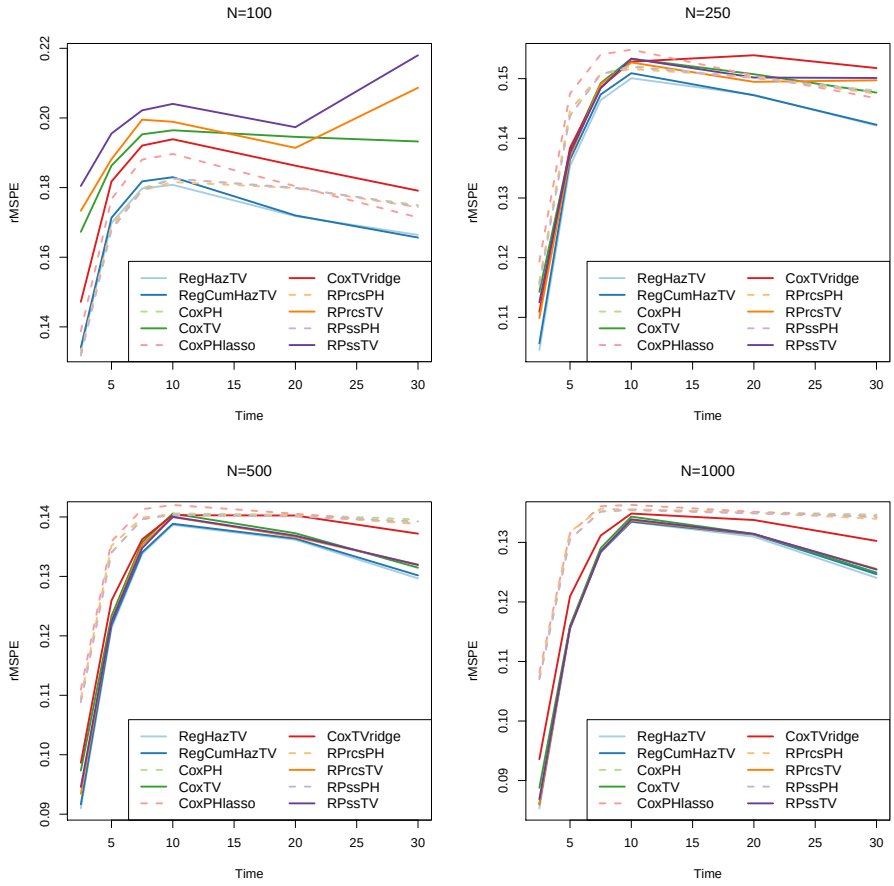


Figure 3.2: Root mean squared prediction error over time each of the evaluated methods. Note that the line for CoxPH is hardly visible since its curve is almost identical to the curves of RPrcsPH and RPssPH. Solid lines are for models allowing for time-varying coefficients; dashed lines for proportional hazard models.

size settings. Their benefit was most apparent for later prediction times. As for the time-averaged results, the proportional hazards methods were at a clear

disadvantage in large sample size settings (due to misspecification), but had the edge over non-regularized TV methods in smallest sample size setting (due to decreased risk of overfitting). This pattern was not clearly time-dependent.

The supplementary material shows discriminative performance (Part S3.3) and calibration performance (Part S3.4) in the validation data. Results were in line with the rMSPE results, with RegHazTV and RegCumHazTV consistently performing well in terms of both time-averaged discriminative performance and discriminative performance at each of the fixed times points. With respect to calibration, average calibration curves across simulations converged towards the diagonal with increasing sample size for all time-varying effects methods except for CoxTVridge. Average calibration curves for RegHazTV and RegCumHazTV did so faster than other time-varying effects methods. PH-method curves clearly reflected misspecification in the larger sample size settings, especially for early and late time-points. The latter was according to expectation since the trends of the true time-varying effects over time were all monotone and thereby cross the time-constant (PH) approximation that captures the average across time. This results in good approximations near the crossing point and bad approximations further away, with the bad approximations corresponding to (on average) high survival probabilities (early in time) and low survival probabilities (later in time).

3.8 Veterans' Administration Lung Cancer study

The Veterans' Administration Lung Cancer (VALC) study is a randomized trial of two chemotherapy treatments in males with advanced inoperable lung cancer [66]. The primary endpoint was time to death and 128 out of 137 patients died during follow-up (the remainder being censored). Data on a selection of variables is available in Kalbfleisch and Prentice [66] and includes time-to-event, event status, and baseline data on treatment (standard vs new chemotherapy), age (in years), prior therapy (yes/no), histological type (squamous, small cell, adeno, large cell), performance status (Karnofsky rating from 0-100, with higher scores relating to better status), and time between diagnosis and randomization (in months). A Cox proportional hazards model including all of these measures as main effects shows clear signs of non-proportionality based on the Grambsch and Therneau test on Schoenfeld residuals [75] ($p = 3.2e^{-5}$), with clear individual contributions of cell type ($\chi^2_3 = 15.2, p = 0.0016$) and Karnofsky rating ($\chi^2_3 = 12.9, p = 0.0003$). This provides us with an interesting setting to illustrate all of

the methods used in the simulation study. The models and methods as applied in the VALC data were as listed below.

1. Regularized log hazard model including time-varying effects (RegHazTV), with the log baseline modeled with a restricted cubic spline with 4 degrees of freedom, all main effects, and including linear interactions with log time. The log baseline hazard parameters were penalized with a ridge penalty, and the remaining parameters with a lasso penalty (group lasso in case of cell type which had 3 groups).

2. Regularized log cumulative hazard model including time-varying effects (RegCumHazTV): same as (1), but on the log cumulative hazard scale.

3. Cox proportional hazards model (CoxPH): a Cox proportional hazards model with all main effects. Predicted survival was derived based on the Cox model and the corresponding Breslow estimate of the cumulative baseline hazard.

4. Cox model including time-varying effects (CoxTV): same as (3), but allowing for time-varying effects as a linear function of log time.

5. Cox proportional hazards lasso model (CoxPHridge): same as (3), but with a ridge penalty on all parameters. Ridge was preferred over lasso due to presence of a categorical variable with 3 groups (cell type).

6. Cox time-varying effects ridge model (CoxTVridge): same as (4), but with a ridge penalty on all parameters.

7. Royston-Parmar proportional hazards model (RPrcsPH): a proportional Royston-Parmar model (*i.e.* log cumulative hazard model) with a 4 degrees of freedom natural cubic spline for the log cumulative baseline hazard and all main covariate effects.

8. Royston-Parmar time-varying effects model (RPrcsTV): same as (7), but including interactions with log time for all ten covariates.

9. Royston-Parmar proportional hazards model (RPssPH): a proportional Royston-Parmar
model (*i.e.* log cumulative hazard model) with a smoothing spline for the log cumulative baseline hazard and all main covariate effects.

10. Royston-Parmar time-varying effects model (RPssTV): same as (9), but including interactions with log time for all ten covariates.

11. Regularized log hazard model (RegHazPH), same as (1), but without the time-varying effects.

12. Regularized log cumulative hazard model (RegCumHazPH): same as (2), but without the time-varying effects.

Due to the limited sample size in the VALC data, note that, compared to the simulation study, 1 df less was spent on the baseline hazard for parametric models, and that time-varying effects were modeled linearly instead of using splines for all methods where applicable. For the same reason, the last two models were added as simplifications of the first two in light of the simulation results.

A bootstrapping approach was used to evaluate model performance. All penalty parameters were selected based on 10-fold cross-validation as performed in (*i.e.* nested in) bootstrap samples. Performance measures were derived in out-of-bag samples. Performance was measured in terms of time-dependent Brier score [77], time dependent c-statistic [78] and graphical calibrations curves [76]. The time dependent c-statistic as described by Antolini et al. [78] was adapted to match Harrel's C [79] in case of proportional hazards by counting tied prediction for discordant outcomes as 0.5 instead of 0. Time-points for the derivation of an integrated Brier score were the $.1, .2, \ldots, .9$ quantiles of the event times distribution in the full data set (with equal weights for all time-points). Graphical calibration curves were derived at the median event time (t=62). A total of 500 bootstrap runs was performed.

Results are shown in Table 3.1 and Figure 3.3 for all methods except CoxTVridge, which did not converge regardless of the choice of penalty. Even though the differences were small, RegHazTV, RegCumHazTV, RegHazPH, RegCumHazPH, and RPrcsTV performed significantly better than the remaining methods in terms of squared prediction error. In terms of rank-ordering, the two regularized proportional hazards models performed best (RegHazPH, RegCumHazPH). Figure 3.3 shows calibration performance at median event time (t=62). RegHazPH and RegCumHazPH again performed well and the curves for the other regularized models also look reasonable and in agreement with the Brier scores. Summarizing, the proportional regularized parametric models performed best on all measures, but differences were small. Even though allowing for time-varying effects is quite a stretch given the limited sample size, the regularized time-dependent effects models performed reasonably well.

	Briertd (se)	**Ctd (se)**
RegHazTV	0.152 (0.042)	0.698 (0.044)
RegCumHazTV	0.152 (0.042)	0.695 (0.049)
CoxPH	0.155 (0.043)	0.705 (0.036)
CoxTV	0.154 (0.043)	0.682 (0.057)
CoxPHridge	0.154 (0.043)	0.708 (0.035)
RPrcsPH	0.154 (0.043)	0.705 (0.036)
RPrcsTV	0.152 (0.043)	0.692 (0.044)
RPssPH	0.154 (0.043)	0.705 (0.036)
RPssTV	0.153 (0.042)	0.676 (0.060)
RegHazPH	0.152 (0.042)	0.713 (0.035)
RegCumHazPH	0.152 (0.042)	0.713 (0.035)

Table 3.1: Mean and standard error of the integrated time-dependent Brier score (Briertd) and the time-dependent c-statistics (Ctd) are shown as derived based on 500 out-of-bag estimates for the Veterans' Administration Lung Cancer study.

3.9 Software

Regularized log hazard and log cumulative hazard modeling has been implemented in `R` [60] package **regsurv**. A development version of the package is available on GitHub `https://github.com/jeroenhoogland/regsurv` and provides functions for model optimization and penalty parameter tuning, as well as convenience functions for prediction and plots of loss and coefficients paths across a penalty parameter grid. Royston-Parmar modeling software is readily available (*e.g.* **stpm2** [57] in `Stata` [58] and **rstpm2** [59] in `R`, and the same holds for standard cox modeling (*e.g.* the **survival** package [80] in `R`) and regularized cox modeling (*e.g.* the **glmnet** package [81] in `R`). R script for replication of the simulation study and applied example is available as supplementary material.

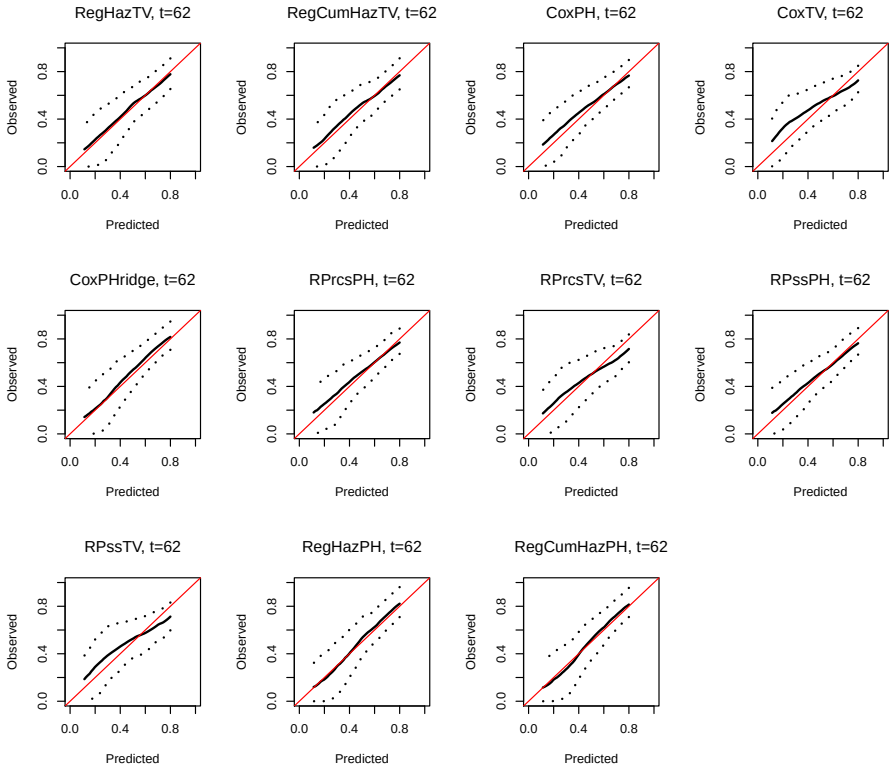


Figure 3.3: Calibration curves for out-of-bag predictions at median event time in the VALC data. Solid lines represent the average calibration curve over 500 out-of-bag estimates; dotted lines are for the 10^{th} and 90^{th} percentile.

3.10 Discussion

We have introduced regularization methods for parametric survival models with a flexible baseline hazard or cumulative hazard. This opens an important toolbox that constrains the risk of overfitting and increases prediction accuracy for a flexible class of models. Importantly, these models explicitly model the baseline (cumulative) hazard, which is of interest for absolute risk prediction over time and when modeling time-varying effects.

From a theoretical perspective, it was shown that the optimization problems for the proposed methods can be formulated as convex optimization problems. This entails that there exists a global optimal value. From a pragmatic perspective, convexity allows for the use of well-known solving methods to find a solution corresponding to the global optimal value. The introduced penalty functions include the well-known elastic net penalty (and hence ridge and lasso penalty) and the group lasso penalty. Simulation work showed that the proposed methods performed well in comparison to alternative methods including Cox regression, regularized Cox regression, and Royston-Parmar models of various types. Importantly, regularization was beneficial even in large sample size settings. In line with the simulation results, the applied example in the Veterans' Administration Lung cancer study showed that the proposed methods performed well in terms of squared prediction error, rank-ordering, and calibration. A software implementation for model fitting, penalty parameter tuning, and prediction has been developed in the form of `R` package **regsurv** and is available on GitHub.

Regarding practical implementation, the choice between log hazard and log cumulative hazard modeling deserves some further attention. The log cumulative hazard is naturally constrained to be monotone non-decreasing (*i.e.* non-negative hazards). Both our implementation and the unregularized Royston-Parmar implementation by Liu et al. [59] enforce this constraint in the development data, while this is not the case in the original proposal [52]. Nonetheless, monotonicity depends on the data in case of time-varying effects models and can hence not be guaranteed when the model is applied in new data. In our experience, such predictions of negative hazards (at some time points for some patients) occur more frequently for models with many time-varying components. While the final contribution of such a time point with a negative predicted hazard to a predicted probability may be small, it is inconvenient. Possible solutions are not straightforward and possibly cumbersome; one idea is to limit model applicability to a multivariable domain in the predictor space where the model always satisfies the monotonicity constraint. While modeling of the log hazard is computationally inconvenient, it does provide an unconstrained optimization problem and a model that generalizes to new data without problem. Therefore, we prefer log hazard models when incorporating time-varying effects.

With respect to limitations, it should be noted that the simulation study and applied example were intended as a proof-of-concept for the introduced methodology, and future research is needed to investigate performance in a wider range of settings. Also, the computation time for regularization paths can be considerable for the log hazard models due to the required numerical integration. For

example, individual solving times for fixed tuning parameters took up to 10 seconds for the large sample size setting for the log hazard models, whereas these took only around half a second for the log cumulative hazard models. Nonetheless, computation times for the **glmnet** implementation of Cox regression with time-varying effects may be even longer without coarsening the grid of event times used to represent time-varying effects.

The current application of convex optimization techniques in the context of parametric survival modeling is a particular example of a widely applicable technique [71, 72]. Any penalty that can be formulated as a convex function can be added to any log-likelihood that can be formulated as a convex function, with the option to include convex constraints. The recent implementation of an elastic net penalty for the family of transformation models provides a recent example illustrating the generality of the approach [82]. The challenge is in the recognition of particular problems as convex problems.

As final note on the use of penalized maximum likelihood, its application in the context of regularization of parameter size (as in the current work) should be distinguished from its application in the context of penalized splines. The first shrinks model parameters towards zero, while the latter shrinks functional forms to linear (*e.g.* by penalizing curvature as quantified by second derivatives). Examples of the latter in the context of parametric survival modeling include the earlier mentioned work of Liu et al.[59] that uses the generalized additive modeling framework [83] to enable flexible modeling of the log cumulative baseline hazard (*e.g.* using smoothing splines). A similar synergy between generalized additive modeling and log hazard models has recently been implemented by Fauvernier et al. [84] (available in `R` package `penSurv` [85]). Ideally, penalized splines are used to shrink functional form to linear, and then ridge/lasso is used to shrink further if needed. Such a hybrid approach has already been described and implemented in R package `gamsel` for gaussian and binomial families [86].

Summarizing, parametric log hazard and log cumulative hazard models provide a flexible tool for survival analysis, and the current addition of regularization further enhances flexibility while controlling for overfitting in settings with limited sample size in light of model complexity. This is of particular interest for the development of prediction models with the aim to predict survival probabilities over time.

Acknowledgements

J Hoogland, TPA Debray, J IntHout, and JB Reitsma acknowledge financial support from the Netherlands Organisation for Health Research and Development (grant 91215058). TPA Debray also acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050)

Data Availability Statement

Data for the Veterans' Administration Lung cancer study are publicly available in the **survival** package [80] in `R` or in Kalbfleisch and Prentice [66]. `R` scripts to perform the simulation study, including data generation and analysis, are available for sharing.

Supplementary Material

The supplementary material consists of further information on restricted cubic splines (Part S3.1), the data generating mechanism used in the simulation study (Part S3.2), and discrimination and calibration results for the simulation study (Part S3.3 and Part S3.4 respectively).

S3.1 Restricted cubic spline details

For set of knots $\boldsymbol{k} = k_1, \ldots, k_m$, including outer knots taken to be the lower (k_1) and upper (k_m) limit of the observed range of u, and $m-2$ inner knots equal ordered quantiles of u, restricted cubic spline $s(u|\boldsymbol{\alpha}, \boldsymbol{k})$ is defined as [67]

$$s(u|\boldsymbol{\alpha}, \boldsymbol{k}) = \alpha_0 v_0 + \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m = \boldsymbol{V}\boldsymbol{\alpha} \tag{3.17}$$

where v_0 is a vector of 1s, $v_1 = u$, and

$$v_{j+1} = \frac{1}{\tau}\left\{(u-k_j)_+^3 - \frac{k_m - k_j}{k_m - k_{m-1}}(u-k_{m-1})_+^3 + \frac{k_{m-1} - k_j}{k_m - k_{m-1}}(u-k_m)_+^3\right\}$$

for $j = 1, \ldots, m$ and $\tau = (k_m - k_1)^2$.

Derivative with respect to t In our implementation of the log cumulative hazard models, $u = \ln(t)$ and the derivative of $s(u|\boldsymbol{\alpha}, \boldsymbol{k})$ with respect to t is required for computation of the hazard. This derivative can be written as

$$s'(u|\boldsymbol{\alpha}, \boldsymbol{k}) = \alpha_0 v_0' + \alpha_1 v_1' + \alpha_2 v_2' + \ldots + \alpha_m v_m' = \boldsymbol{V'\alpha} \tag{3.18}$$

where v_0' is a vector of 0s, v_1' is a column of $1s$, and

$$v_{j+1}' = \frac{3}{t}\frac{1}{\tau}\left\{(u-k_j)_+^2 - \frac{k_m - k_j}{k_m - k_{m-1}}(u-k_{m-1})_+^2 + \frac{k_{m-1}-k_j}{k_m - k_{m-1}}(u-k_m)_+^2\right\}$$

for $j = 1, \ldots, m$ and $\tau = (k_m - k_1)^2$. Note that only the equation for the basis functions differs from $s(\cdot)$ and that $\boldsymbol{\alpha}, \boldsymbol{k}$, τ, and the scaling factors $\frac{k_m - k_j}{k_m - k_{m-1}}$ and $\frac{k_{m-1}-k_j}{k_m - k_{m-1}}$ are exactly equivalent between $s'(\cdot)$ and $s(\cdot)$.

Non-proportional hazards The covariate $\times$ log time interactions for a single covariate column $\boldsymbol{z_j}$, knots $\boldsymbol{\kappa}$ and coefficients $\boldsymbol{\gamma}$ is defined as

$$s(u, \boldsymbol{z_j}|\boldsymbol{\gamma}, \boldsymbol{\kappa}) = \boldsymbol{DV\gamma}$$

where $\boldsymbol{D}_{n\times n} = \text{diag}(\boldsymbol{z_j})$, $\boldsymbol{V}_{n\times m}$ contains the basis columns $v_1, \ldots, v_m$ for u and knots $\boldsymbol{\kappa}$, and $\boldsymbol{\gamma}$ is a column vector with the corresponding coefficients. We use $s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ to denote the concatenation of $s(u, \boldsymbol{z}_j|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ for each of the covariates in set I to be modelled as non-proportional. That is, for covariates $1, \ldots, q$ in set I,

$$s(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa}) = \boldsymbol{D_1 V\gamma_1} + \ldots + \boldsymbol{D_j V\gamma_j} + \ldots + \boldsymbol{D_q V\gamma_q}$$

where $\boldsymbol{D_j}$ is a diagonal matrix with the j'th non-proportional covariate on the diagonal, $\boldsymbol{V}$ is as above and is therefore constant over j, and $\boldsymbol{\gamma_j}$ contains the coefficients for the interactions with covariate $\boldsymbol{z}_j$. The derivatives $s'(u, \boldsymbol{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ with respect to t can be derived analogously to the main effect restricted cubic splines and only involve substitution of the matrix $\boldsymbol{V}$ by $\boldsymbol{V'}$ containing the $\frac{\partial}{\partial t}$ columns of $\boldsymbol{V}$.

S3.2 Data generating mechanism

Figure S3.1 shows baseline survival, cumulative baseline hazard, baseline hazard, and the time-varying effects corresponding to the data generating mechanism used in the simulation study.

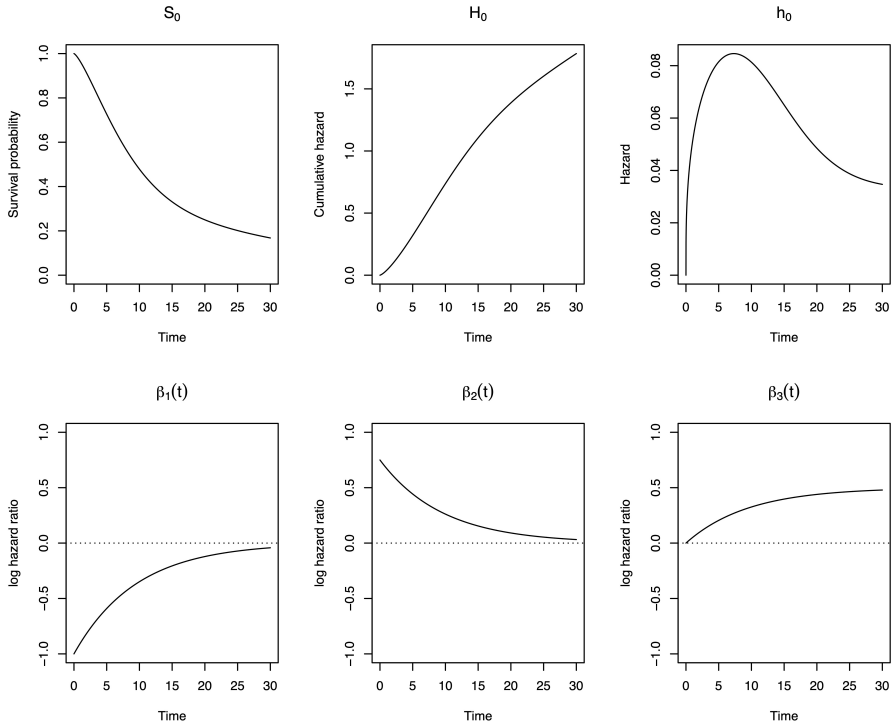
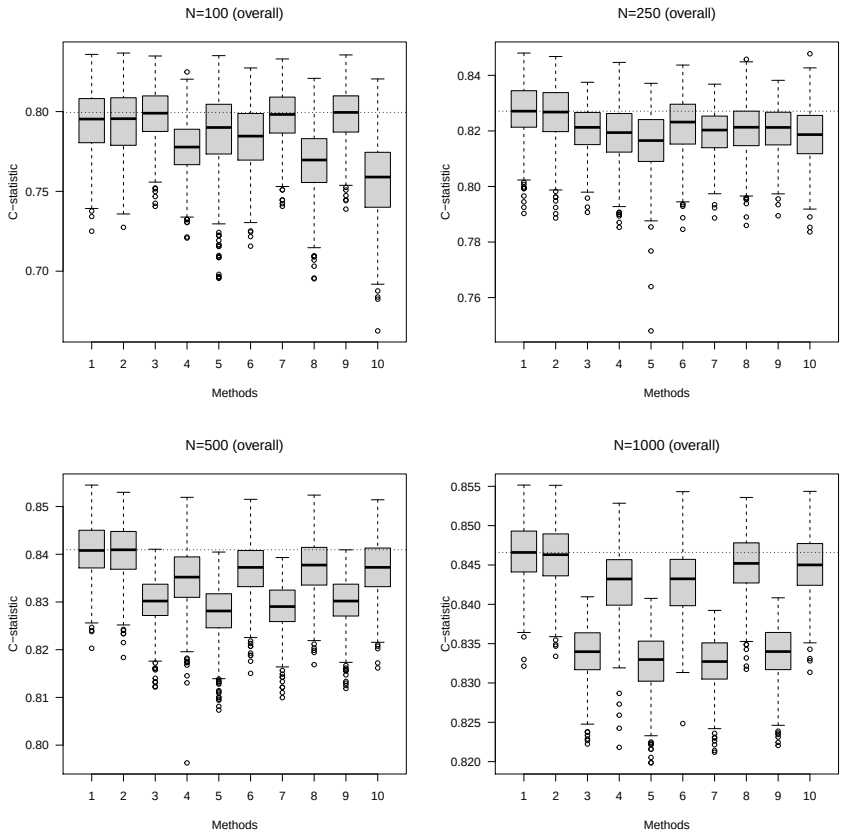
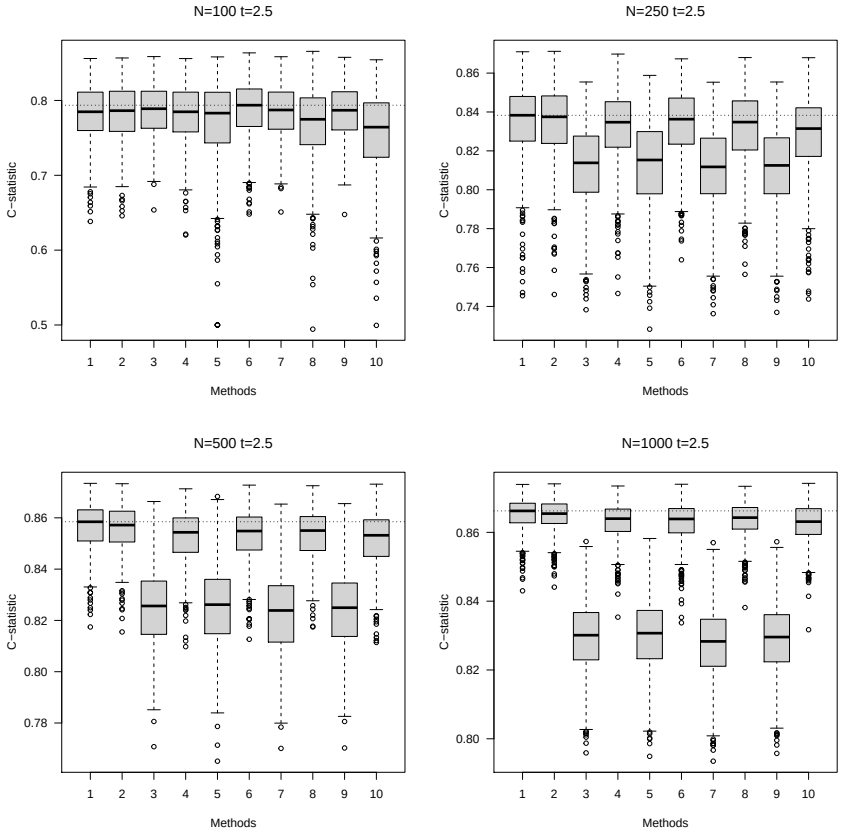


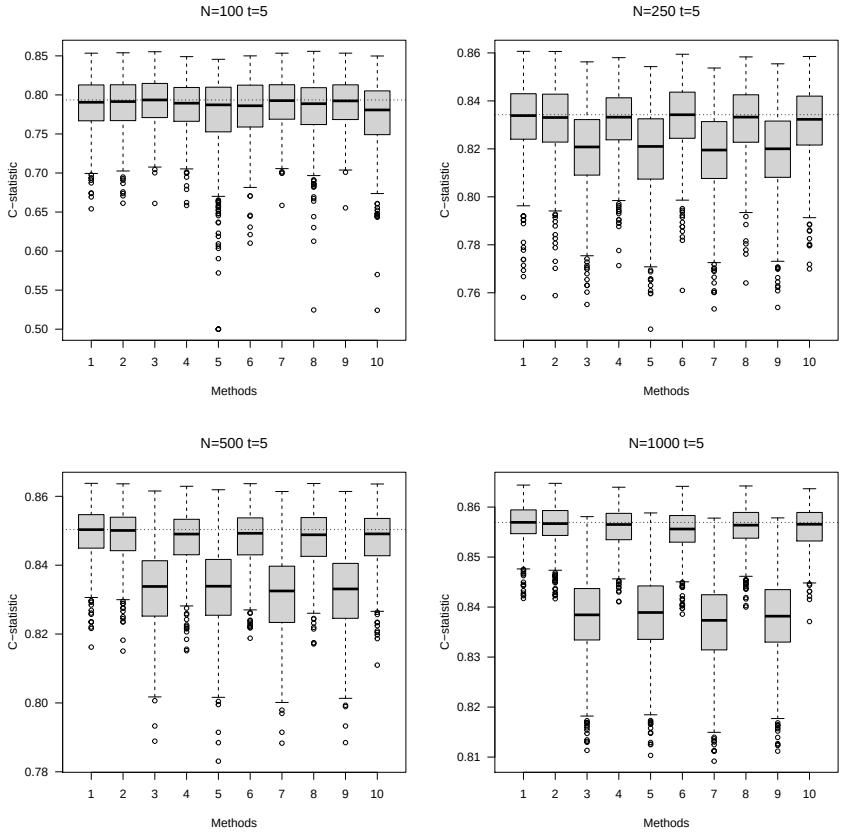
Figure S3.1: Baseline survival (top left), cumulative baseline hazard (top middle), baseline hazard (top right), and the time-varying effects for the first three covariates (bottom row) for the data generating mechanism used in the simulation study.

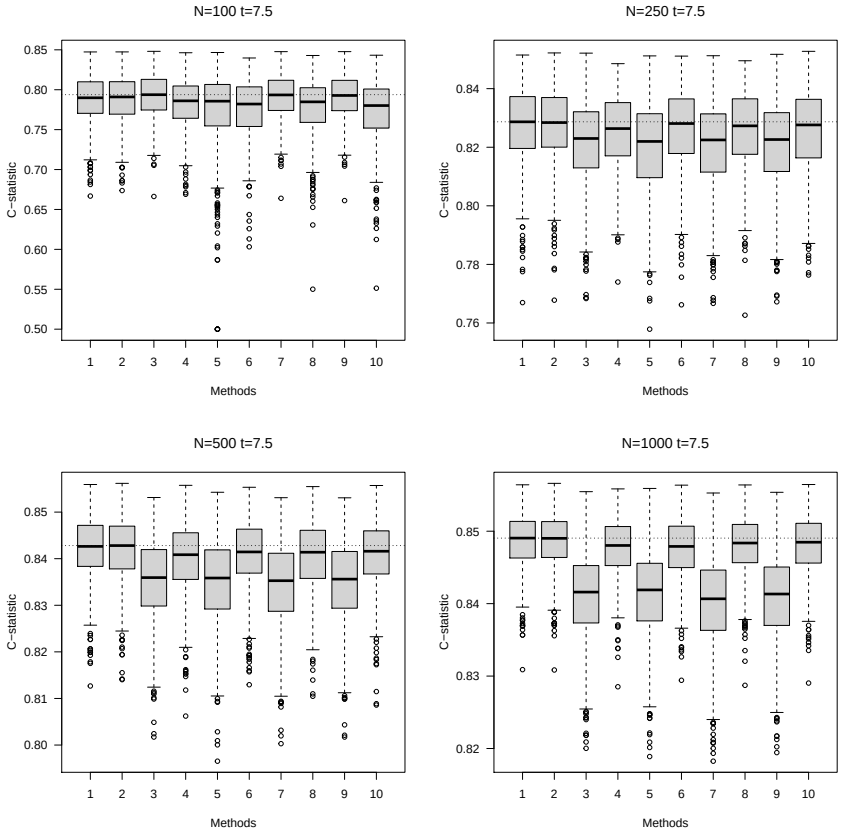
S3.3 Discrimination results

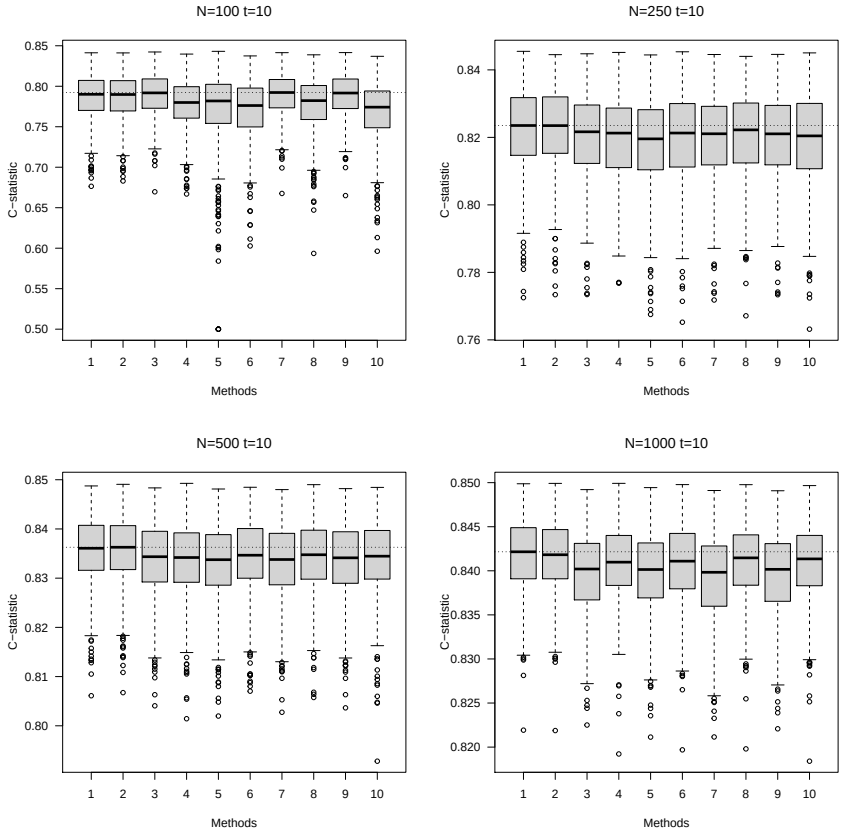
Each of the figures shows C-statistics for each of the 10 methods under evaluation: (1) RegHazTV, (2) RegCumHazTV, (3) CoxPH, (4) CoxTV, (5) CoxPHlasso, (6) CoxTVridge, (7) RPrcsPH, (8) RPrcsTV, (9) RPssPH, (10) RPssTV. Separate figures are provided for each sample size setting and for time-averaged C-statistics and fixed time-point C-statistics. Boxplots reflect results from 500 simulation runs each.

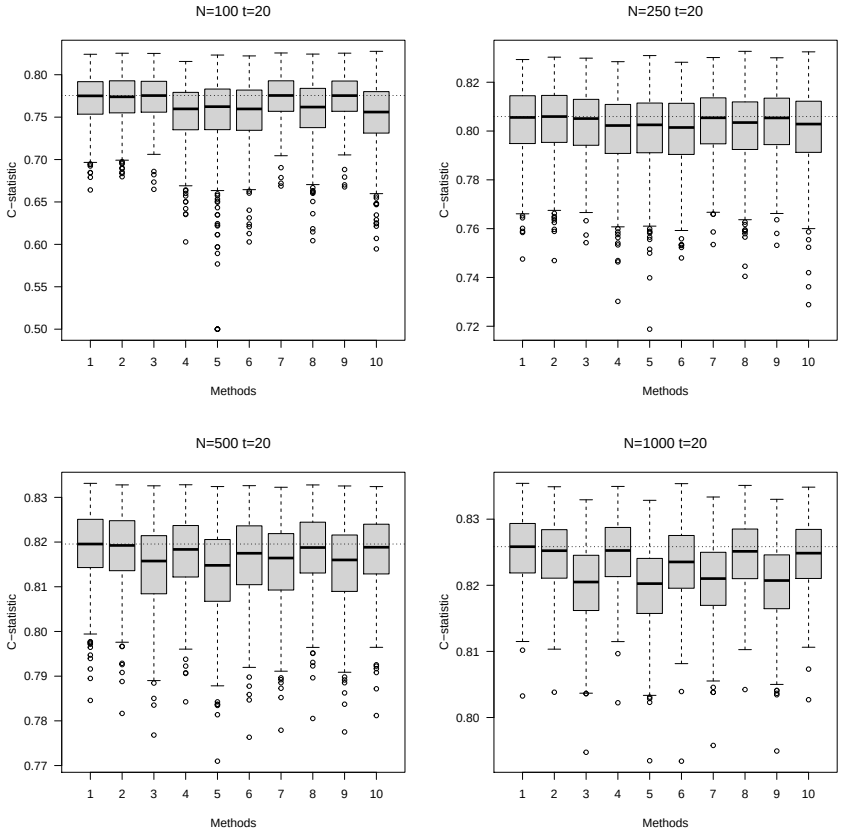

N=100 (overall)
N=250 (overall)
N=500 (overall)
N=1000 (overall)
C−statistic
Methods
1
2
3
4
5
6
7
8
9
10

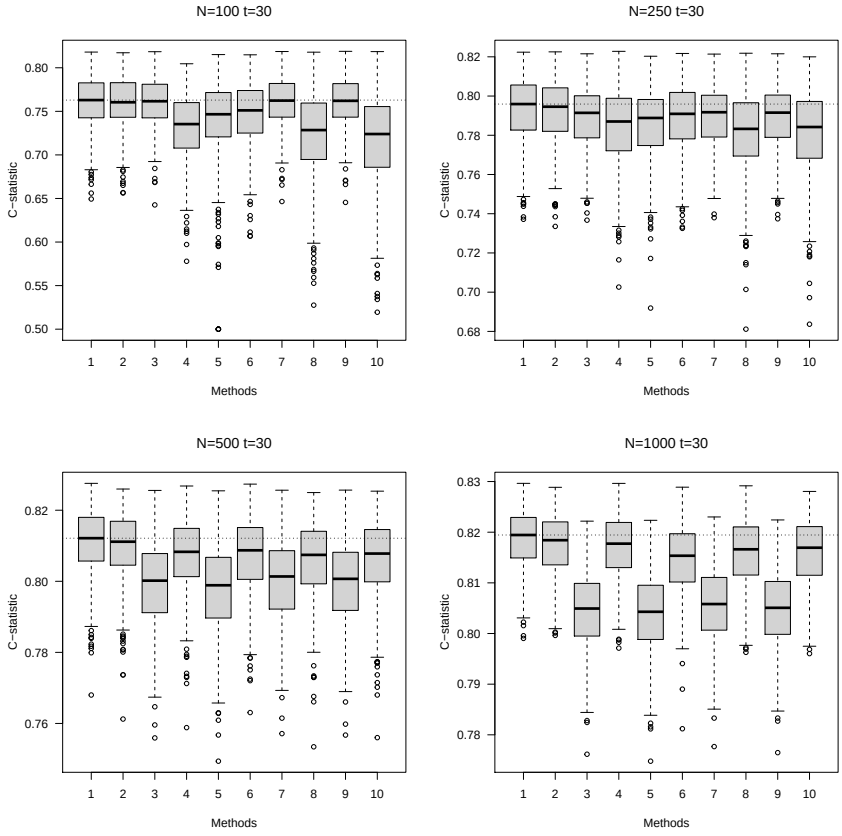

N=100 t=2.5
N=250 t=2.5
N=500 t=2.5
N=1000 t=2.5
C−statistic
Methods


N=100 t=5
N=250 t=5
N=500 t=5
N=1000 t=5
C−statistic
Methods


N=100 t=7.5
N=250 t=7.5
N=500 t=7.5
N=1000 t=7.5
C−statistic
Methods

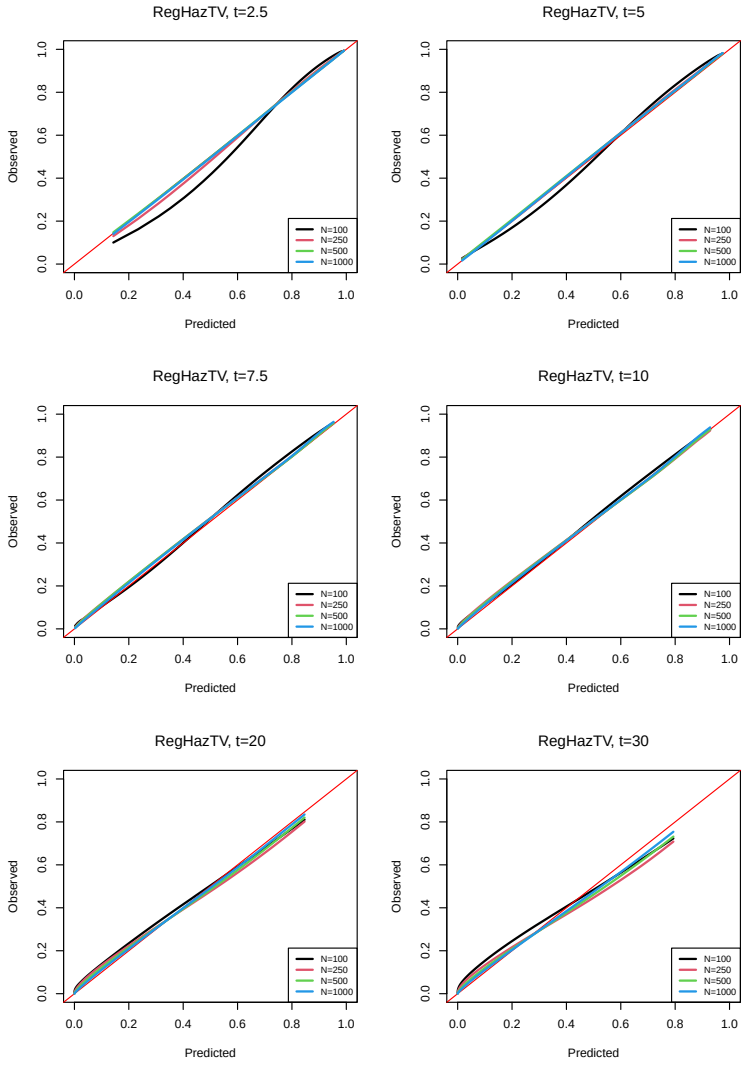

N=100 t=10
N=250 t=10
N=500 t=10
N=1000 t=10
C−statistic
Methods

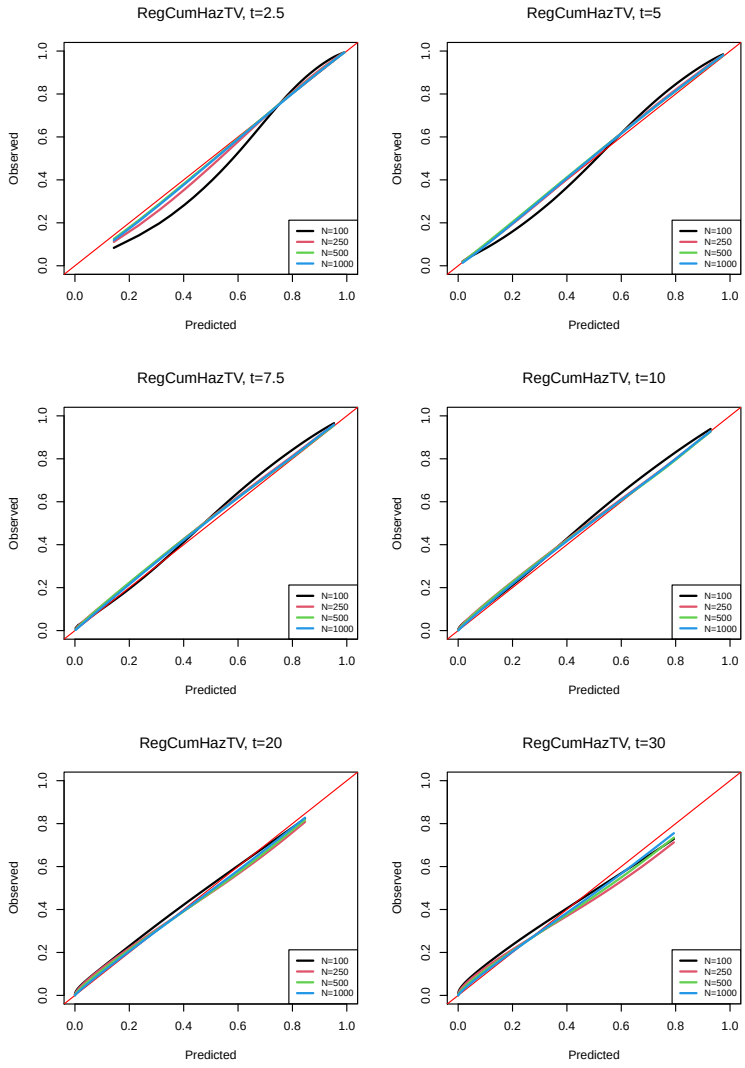

N=100 t=20
N=250 t=20
N=500 t=20
N=1000 t=20
C−statistic
Methods

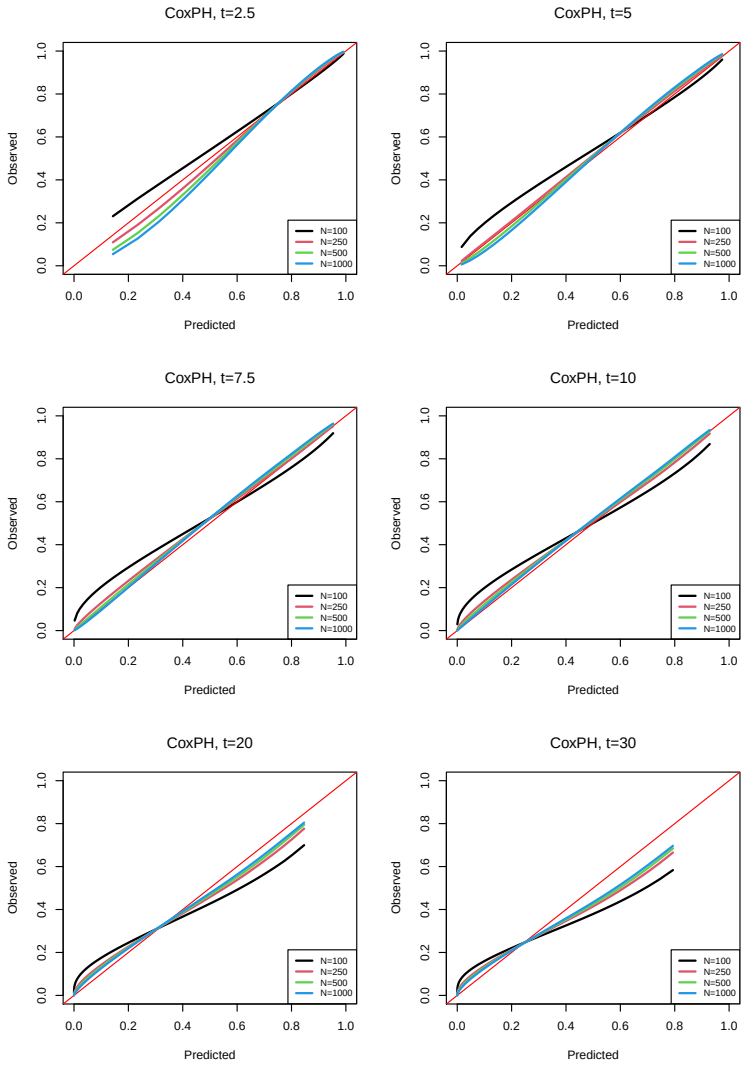

N=100 t=30
N=250 t=30
N=500 t=30
N=1000 t=30
C−statistic
Methods

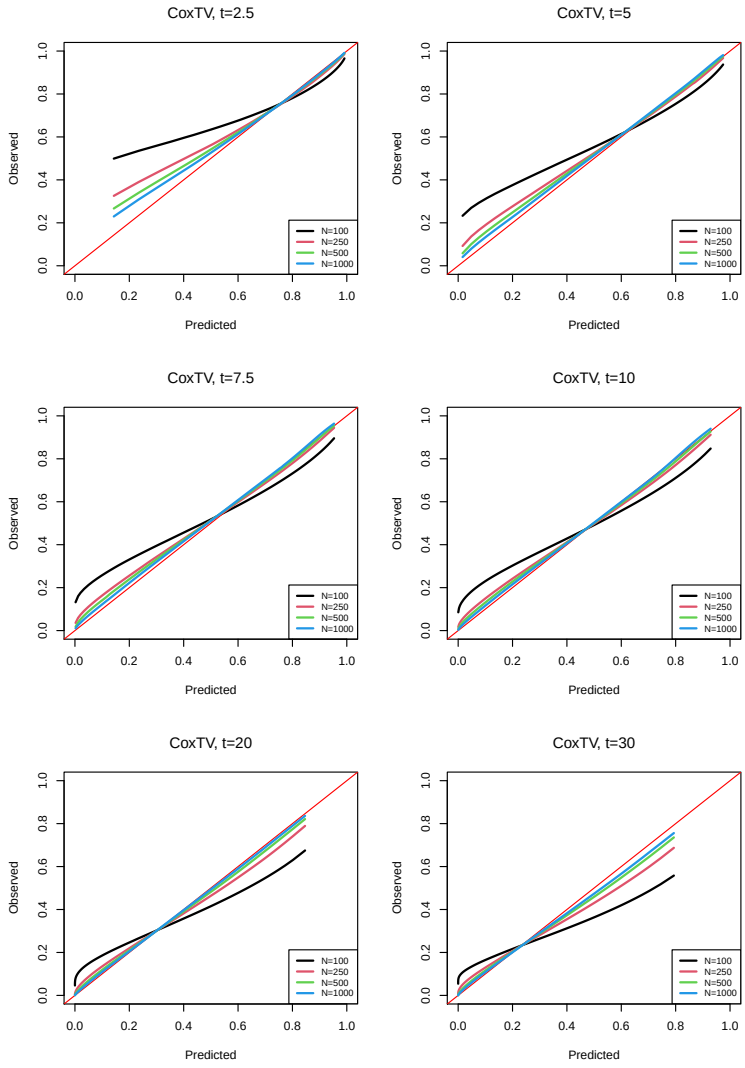
S3.4 Calibration results

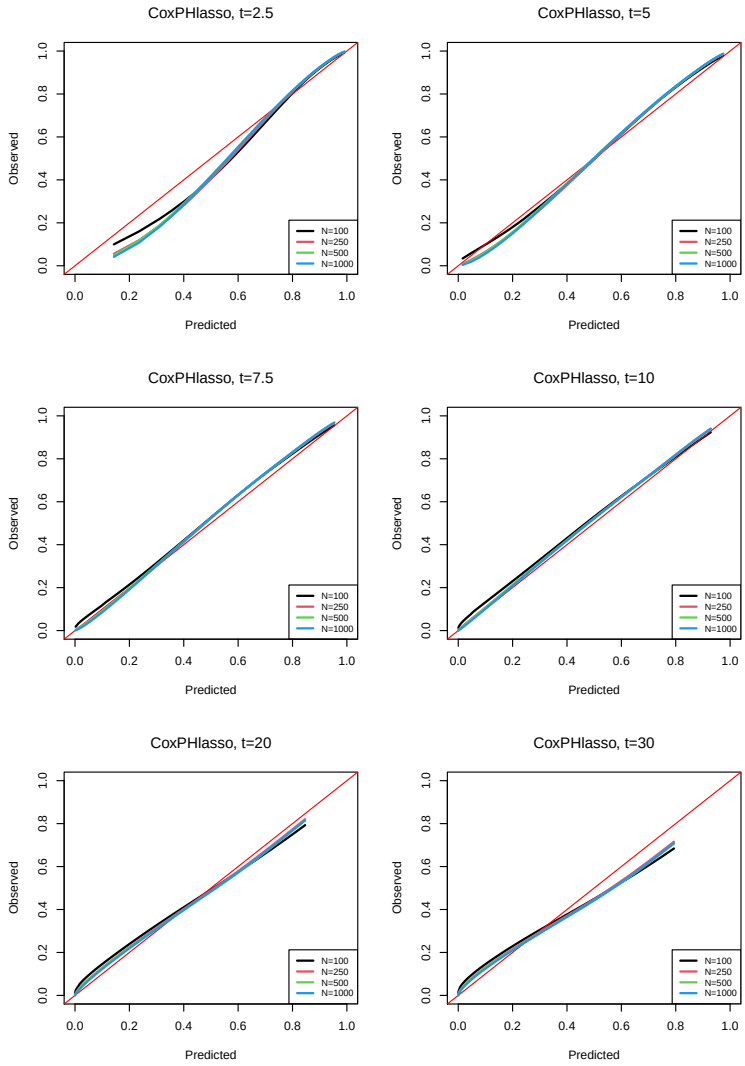
Each of the figures shows calibration curves as averaged over simulation runs for each of the sample size settings, time-points, and methods under evaluation.

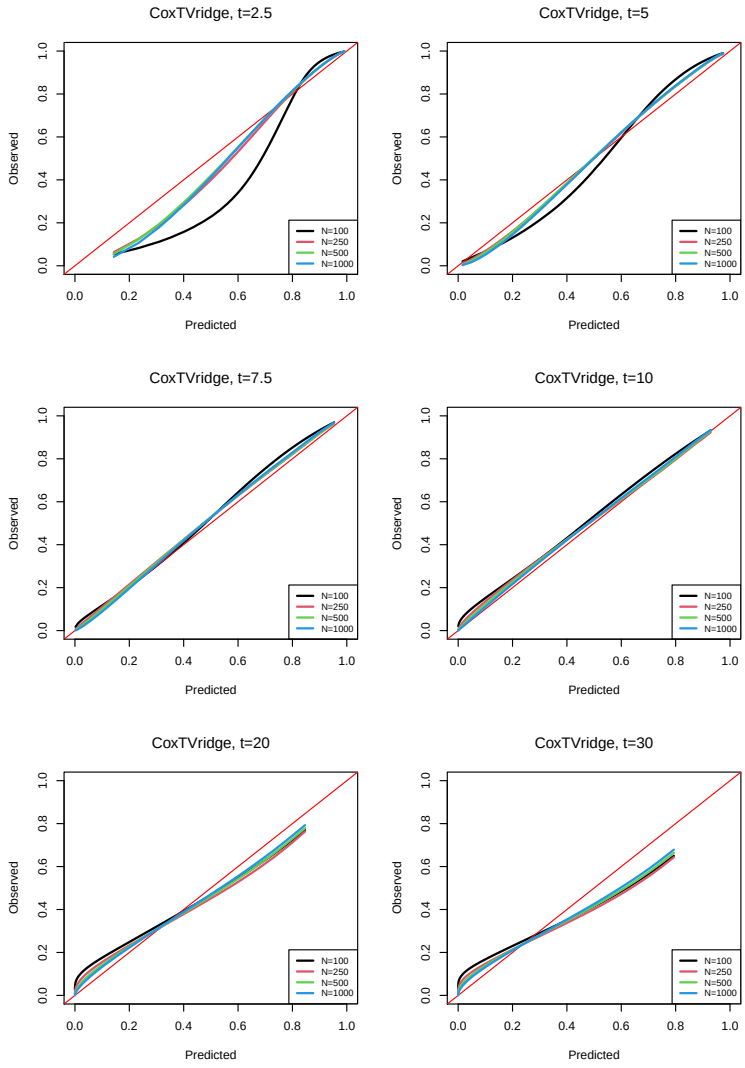


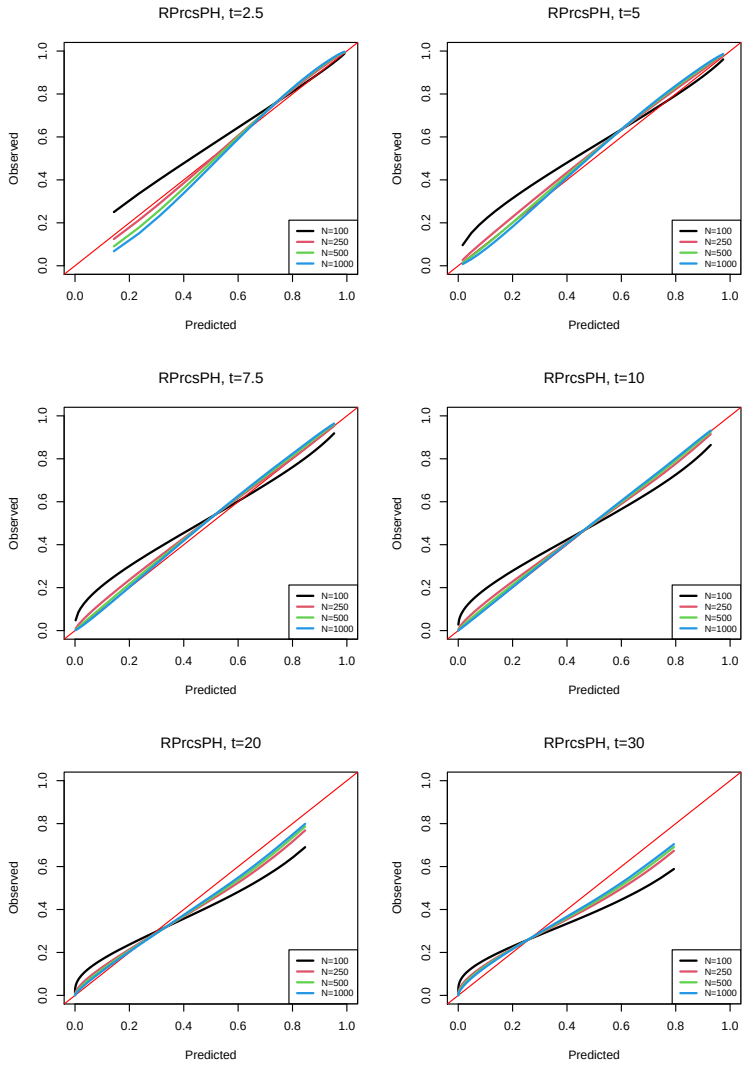

RegCumHazTV, t=2.5
RegCumHazTV, t=5
RegCumHazTV, t=7.5
RegCumHazTV, t=10
RegCumHazTV, t=20
RegCumHazTV, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000

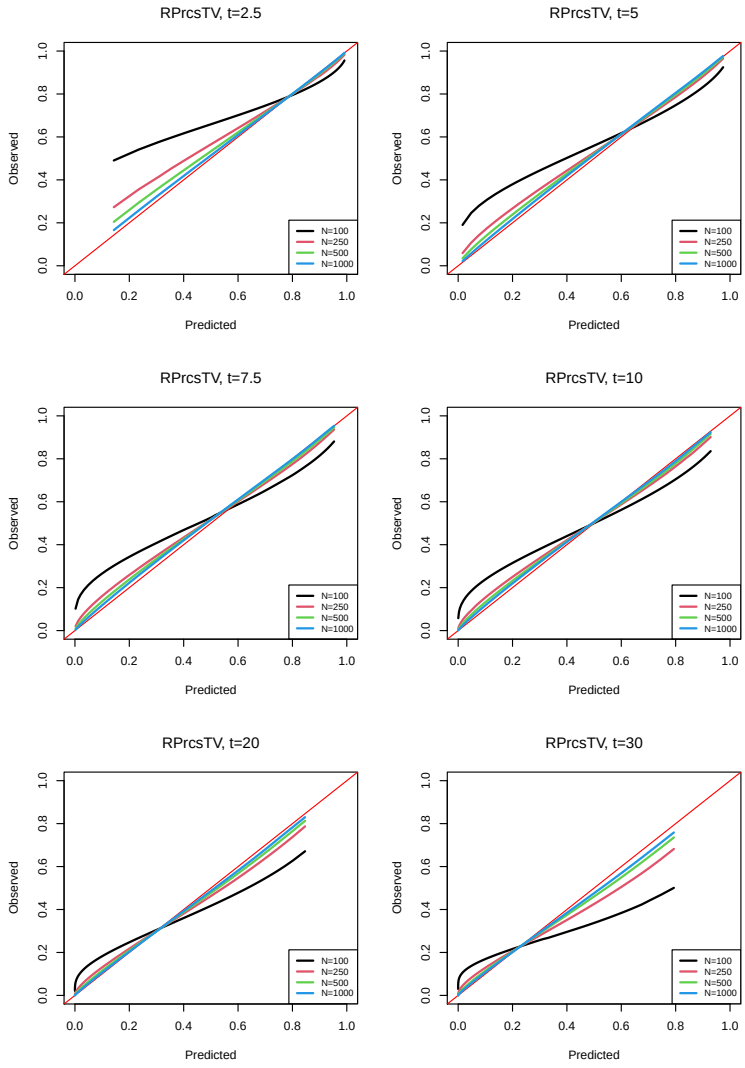

CoxPH, t=2.5
CoxPH, t=5
CoxPH, t=7.5
CoxPH, t=10
CoxPH, t=20
CoxPH, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000

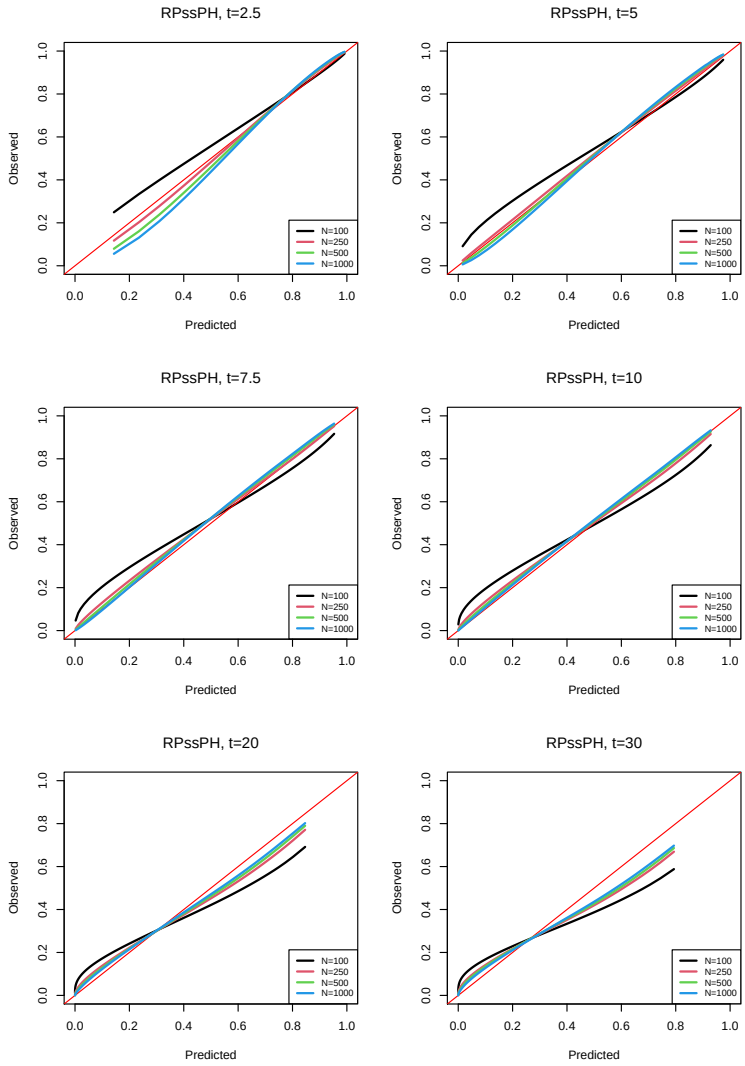

CoxTV, t=2.5
CoxTV, t=5
CoxTV, t=7.5
CoxTV, t=10
CoxTV, t=20
CoxTV, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000

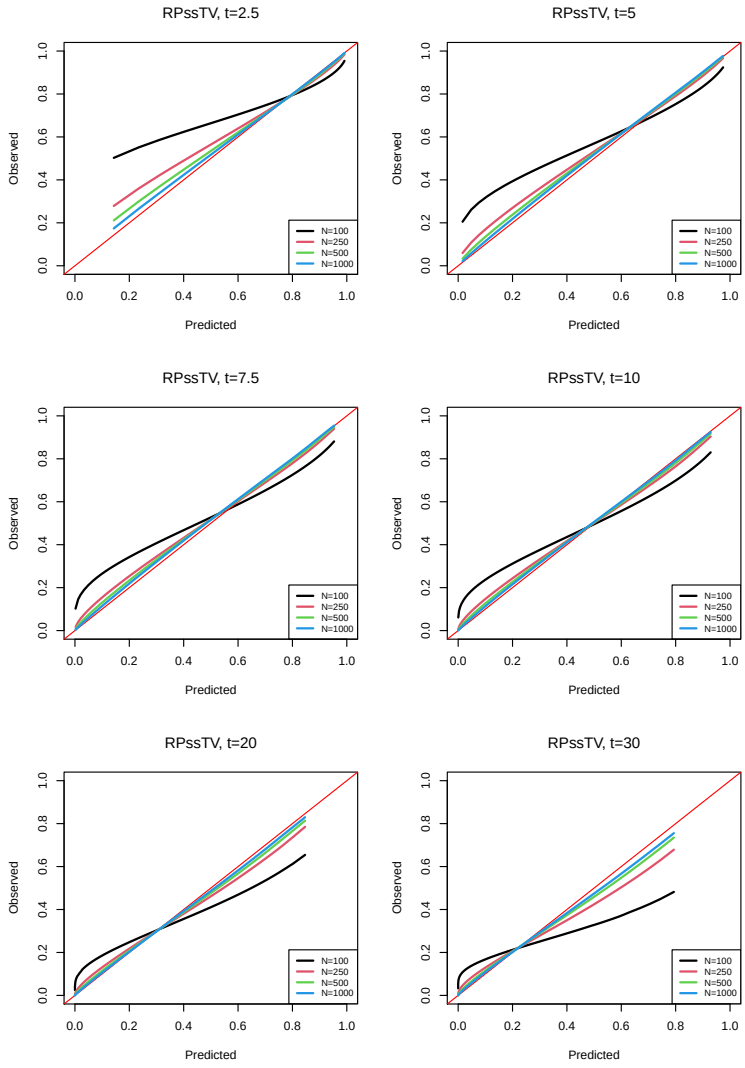

CoxPHlasso, t=2.5
CoxPHlasso, t=5
CoxPHlasso, t=7.5
CoxPHlasso, t=10
CoxPHlasso, t=20
CoxPHlasso, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000


CoxTVridge, t=2.5
CoxTVridge, t=5
CoxTVridge, t=7.5
CoxTVridge, t=10
CoxTVridge, t=20
CoxTVridge, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000


RPrcsPH, t=2.5
RPrcsPH, t=5
RPrcsPH, t=7.5
RPrcsPH, t=10
RPrcsPH, t=20
RPrcsPH, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000


RPrcsTV, t=2.5
RPrcsTV, t=5
RPrcsTV, t=7.5
RPrcsTV, t=10
RPrcsTV, t=20
RPrcsTV, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000


RPssPH, t=2.5
RPssPH, t=5
RPssPH, t=7.5
RPssPH, t=10
RPssPH, t=20
RPssPH, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000


RPssTV, t=2.5
RPssTV, t=5
RPssTV, t=7.5
RPssTV, t=10
RPssTV, t=20
RPssTV, t=30
Observed
Predicted
0.0
0.2
0.4
0.6
0.8
1.0
N=100
N=250
N=500
N=1000

Chapter 4

A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint

Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, Harrell Jr FE, Moons KGM, Debray TPA, Reitsma JB. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*. https://doi.org/10.1002/sim.9154

Abstract
Randomized trials typically estimate average relative treatment effects, but decisions on the benefit of a treatment are possibly better informed by more individualized predictions of the absolute treatment effect. In case of a binary outcome, these predictions of absolute individualized treatment effect require knowledge of the individual's risk without treatment and incorporation of a possibly differential treatment effect (i.e. varying with patient characteristics). In this paper we lay out the causal structure of individualized treatment effect in terms of potential outcomes and describe the required assumptions that underlie a causal interpretation of its prediction. Subsequently, we describe regression models and model estimation techniques that can be used to move from average to more individualized treatment effect predictions. We focus mainly on logistic regression-based methods that are both well-known and naturally provide the required probabilistic estimates. We incorporate key components from both causal inference and prediction research to arrive at individualized treatment effect predictions. While the separate components are well known, their successful amalgamation is very much an ongoing field of research. We cut the problem down to its essentials in the setting of a randomized trial, discuss the importance of a clear definition of the estimand of interest, provide insight into the required assumptions, and give guidance with respect to modeling and estimation options. Simulated data illustrates the potential of different modeling options across scenarios that vary both average treatment effect and treatment effect heterogeneity. Two applied examples illustrate individualized treatment effect prediction in randomized trial data.

4.1 Introduction

Prediction of risk and prediction of treatment effect are two key components in modern medicine and personalized healthcare. On the one hand, risk predictions are classically functions of multiple patient characteristics. They include predictions of the risk of having a specific health outcome or condition (diagnosis) or of developing a future health outcome (prognosis). Also, risk predictions vary naturally across patients, are descriptive, and can be uniformly expressed as probabilities [51]. Importantly, risk prediction models are generally descriptive and are *not* intended to reflect the causal mechanism; in particular, included predictor effects in the model are not intended to reflect the extent to which their removal or modification would change an individual's prediction. On the other hand, predictions of treatment effect *do* express an expected difference due to modification of the treatment condition. They have classically been studied on a group level (e.g. treated group versus control group), often assume a constant effect across individuals, have a causal interpretation, and are traditionally expressed using relative effect measures (*e.g.* odds ratio, relative risk, or hazard ratio) [87].

Risk predictions and treatment effect estimation are two important areas of research but have largely developed in separation, leading to an apparent contradiction between methods for prediction and methods for causal inference. However, answers to many important questions need to bridge the divide. For instance, "How will a possible treatment change predicted outcome risk?" or "Is there variability in the effect of this treatment across patients (*i.e.* differential treatment effect)?". These questions involve both the causal effect of treatment on a targeted health outcome and the adequate incorporation of associations with individual patient characteristics.

It is exactly these types of questions that need to be answered to provide more tailored, stratified, personalized, or precision medicine [88, 89, 90]. The limitations of average relative treatment effects have long been recognized and the promise of a more individualized yet evidence-based approach has been enticing. Such a strategy requires focus on heterogeneity between patients and its relation to risk of the outcome of interest and variability in treatment effect. Also, moving towards more individualized estimates inherently means moving to more absolute expressions of variability in risk and treatment effect that are interpretable on the individual level [91, 92]. For example, predictions of the absolute risk of a future event under different treatment conditions provide a natural basis for shared decision-making.

In this tutorial, we aim to give a platform for statisticians and other researchers embarking on the prediction of individualized treatment effect. We focus on the risk of developing a binary outcome or endpoint, and aim to combine the highly conditional nature of typical risk prediction modeling with causal inference about treatment effectiveness. While the problem is well-known, it is very complex and therefore, we will cut it down to its essentials in the setting of a randomized trial, and mainly limit our scope to regression-based methods. We discuss the importance of a clear definition of the estimand of interest, provide insight into the required assumptions, and give guidance with respect to the modeling and estimation options. Key considerations with respect to the choice of modeling and estimation methods are further illustrated in a simulation study and two applied examples.

4.2 Defining individualized treatment effect

The main idea underlying our endeavor is that the effect of treatment may be different for each individual, and that it may be beneficial to personalize or individualize its estimate. In the context of risk prediction, this implies that treatment causes a change in predicted outcome risk that may vary across individuals conditional on their characteristics. In other words, a personalized or individualized treatment effect describes the effect of modifying a treatment condition (*i.e.* setting its value) while controlling for (*i.e.* conditioning on) that individual's characteristics. We restrict our description to settings in which variables besides treatment do not have a causal interpretation, since this nicely aligns with the typical design of a randomized trial. This lack of causal interpretation for the set of variables conditioned on, is typical for classical prediction modeling. While the inner workings of a model that simultaneously describes both causal and mere associative relations may not need to discern between these different roles, they are of importance when interpreting the model.

To that effect, distinguishing between variables that do and do not have a causal interpretation is helpful for a precise definition of the individualized treatment effect of interest. Two common approaches to make this distinction are the $do(\cdot)$ operator introduced by Pearl [21] and the potential outcomes framework popularized by Rubin [93]. The $do(\cdot)$ operator is an operator that describes the effect of setting or modifying a variable to take a certain value (e.g. $P(Y = y|do(X = x))$) and clearly separates this case from classical conditional notation (e.g. $P(Y = y|X = x)$). The potential outcomes framework, as popularized by

Rubin, allows for a formal distinction at the level of the outcomes that arises when the causal variable takes on different values [94, 93]. For instance, if interest is in the causal effect of a treatment, and treatment takes values $a \in \mathcal{A}$, $Y^{A=a}$ denotes the potential outcome for treatment a. In case of a treatment variable that can be set to 0 (control) or 1 (treated), the two potential outcomes are $Y^{a=0}$ and $Y^{a=1}$. The notation easily allows for conditioning, such that the effect δ of treatment on the risk of an event, conditional on covariates $\boldsymbol{X}$, can be written as

$$\delta(\boldsymbol{x}) = P(Y^{a=1} = 1|\boldsymbol{X} = \boldsymbol{x}) - P(Y^{a=0} = 1|\boldsymbol{X} = \boldsymbol{x}) \tag{4.1}$$

where bold face indicates vectors. The same quantity could be written in *do(·)* notation as

$$\delta(\boldsymbol{x}) = P(Y = 1|do(A = 1), \boldsymbol{X} = \boldsymbol{x}) - P(Y = 1|do(A = 0), \boldsymbol{X} = \boldsymbol{x}) \tag{4.2}$$

For our purposes, the differences between these frameworks are not of interest and we adopt the potential outcomes framework throughout the remainder of the paper for reasons of familiarity in statistical research.

A final remark on the nature of 'individualized' or 'personalized' is in place: the estimand of interest is not truly individual, but relates to groups of individuals sharing a covariate pattern. A truly individual treatment effect can never be observed since only one potential outcome can be observed at any time (or equivalently, only one treatment can be assigned). This problem has been referred to as the fundamental problem of causal inference [23]. Acknowledging this, for a dichotomous outcome Y, we define the individualized treatment effect ($\delta(\boldsymbol{x}_i)$) for individual i with covariate vector $\boldsymbol{x_i}$ as

$$\delta(\boldsymbol{x}_i) = P(Y_i^{a=1} = 1|\boldsymbol{X} = \boldsymbol{x_i}) - P(Y_i^{a=0} = 1|\boldsymbol{X} = \boldsymbol{x_i}) \tag{4.3}$$

The individualized treatment effect $\delta(\boldsymbol{x}_i)$ can be interpreted as the expected difference in outcome risk for an individual with covariate values $\boldsymbol{x}_i$ under two different treatment conditions. This definition is easily extended to ≥ 2 treatment conditions, but we will focus on a setting with 2 treatment conditions. While we will focus on $\delta(\boldsymbol{x}_i)$ as our estimand of interest, note that important information is lost when only looking at the difference between potential

outcomes. Therefore, the main task is to predict the conditional risk of both potential outcomes, which provides a more complete picture and directly leads to an estimate of $\delta(\boldsymbol{x}_i)$. Section 4.3 first discusses the assumptions that allow estimation of $\delta(\boldsymbol{x}_i)$ from randomized trial data. Subsequently, sections 4.4 and 4.5 describe specification and estimation of regression models for the purpose of predicting $\delta(\boldsymbol{x}_i)$.

4.3 Identifiability assumptions

The individualized treatment effect specified in equation (4.3) is written as a difference between two potential outcomes. However, in practice only a single potential outcome will be observed for each individual. Identification of $\delta(\boldsymbol{x}_i)$ based on the observed data requires assumptions to supplement the data. The necessary identifiability assumptions are consistency, exchangeability, and positivity. We here shortly introduce the fundamentals as relevant to our setting; excellent introductory [95] and comprehensive texts on causal inference are available elsewhere [22].

Consistency refers to equality between the observed outcome Y and the potential outcome for the actually assigned treatment Y^a. When A takes on value 0 (control) or 1 (treated), this can be expressed as $Y = AY^{a=1}+(1-A)Y^{a=0}$. This assumption holds when the data reflect well-defined treatments. As a counter example, consider a situation in which there is much variability in the active treatment (e.g. different starting times, intensity or dosage, duration) but these are all just labelled $a = 1$: the causal contrast $Y^{a=1} - Y^{a=0}$ is no longer clearly defined. As is clear from equation (4.3), the contrast of interest actually requires consistency conditional on covariates $\boldsymbol{X}$. This extension is trivial if marginal consistency holds.

Exchangeability requires that the potential outcomes are independent of treatment assignment ($Y^a \perp\!\!\!\perp A$ for all a). In other words, the actually assigned treatment does not predict the potential outcome [95]. As an example, consider a two-arm study of a new treatment: in terms of potential outcomes, exchangeability with respect to treatment here implies that it does not matter which arm received the new treatment. Since our interest is in a conditional treatment effect, exchangeability should hold conditionally ($Y^a \perp\!\!\!\perp A|\boldsymbol{X}$ for all a). While this is a challenging assumption to satisfy in general, it holds automatically in the context of a randomized trial. After either marginal randomization (*i.e.* a common probability of treatment for all) or conditional randomization

(*i.e.* with the probability of treatment depending on covariates, also known as stratified randomization), conditional exchangeability holds when conditioning on (at least) the variables used during randomization. It is important to realize that randomization only provides exchangeability at baseline, and the causal contrast $\delta(\boldsymbol{x}_i)$ at that time therefore reflects an intention-to-treat effect. Any conditioning on post-randomization information is no longer protected by randomization and the exchangeability assumption will no longer be guaranteed to hold [96, 97]. For instance, estimation of a per protocol individualized treatment effect would require further assumptions such as absence of any unmeasured confounders and correct specification of all confounders [98]. We here limit our overview to intention-to-treat effects.

Positivity reflects the assumption that each patient should have a non-zero probability of either treatment assignment, which is clearly fulfilled in case of a randomized study.

A final assumption that is often made is the assumption of *no interference*, stating that the potential outcomes for one individual do not depend on treatment assignment of other individuals. While not strictly necessary, the situation quickly grows in complexity without this assumption since the potential outcome definitions would then have to incorporate the dependence on other units [93, 98]. The combination of consistency and no interference is also often referred to as the *stable unit treatment value assumption* (SUTVA) [93].

The definition of $\delta(\boldsymbol{x}_i)$ in equation (4.3) assumes no interference, which follows from the fact that the potential outcome for individual i only depends on the individual's own covariate status and treatment assignment. Further assuming positivity provides a causal interpretation of the treatment effect conditional on covariates $\boldsymbol{x}_i$, with i from $1, \ldots, n$. Finally, consistency and exchangeability are necessary to re-write the estimand in terms of observed variables only:

$$\begin{aligned}
\delta(\boldsymbol{x}_i) &= P(Y_i^{a=1} = 1 | \boldsymbol{X} = \boldsymbol{x_i}) - P(Y_i^{a=0} = 1 | \boldsymbol{X} = \boldsymbol{x_i}) \\
&= P(Y_i^{a=1} = 1 | A = 1, \boldsymbol{X} = \boldsymbol{x_i}) - P(Y_i^{a=0} = 1 | A = 0, \boldsymbol{X} = \boldsymbol{x_i}) \\
&\quad \text{(by exchangeability)} \\
&= P(Y_i = 1 | A = 1, \boldsymbol{X} = \boldsymbol{x_i}) - P(Y_i = 1 | A = 0, \boldsymbol{X} = \boldsymbol{x_i}) \\
&\quad \text{(by consistency)}
\end{aligned} \tag{4.4}$$

In summary, the identifiability assumptions allow $\delta(\boldsymbol{x}_i)$ to be estimated from the observed data. While equation (4.4) essentially allows for fully non-parametric

estimation of $\delta(\boldsymbol{x}_i)$, there is usually insufficient data to do so when interest is in a highly conditional treatment effect (as is the case for an individualized treatment effect). That is, there will not be sufficient cases with $\boldsymbol{X} = \boldsymbol{x}_i$ under both treatments to reliably estimate $\delta(\boldsymbol{x}_i)$. This brings us to the need of a model for $P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)$ to smooth over the gaps in the observed set of all $\boldsymbol{x}_i$ across both treatments.

4.4 Models for the prediction of individualized treatment effect

With the identifiability conditions in place for a causal interpretation of $\delta(\boldsymbol{x}_i)$ as estimated based on the observed trial data only, the remaining problem can be recognized as a typical prediction modeling problem (equation (4.4)). Therefore, well-established modeling techniques can be used to model the required conditional risks. While a vast array of possible prediction modeling techniques is available, we will focus on modeling techniques that have a basis in generalized linear modeling. More specifically, due to the binary outcome, we will focus on methods that have a basis in logistic regression, which has been the mainstay method for clinical prediction models in settings with a binary endpoint [51]. Key features of logistic regression include that it directly provides the probabilistic estimates of interest [99] and has well-known properties. Also, it is a possibly parsimonious model family for the task at hand as explained in the next section.

4.4.1 Homogeneous treatment effect

The aim is to model the observed outcomes $\boldsymbol{Y}$ as a function of treatment assignment A and covariates $\boldsymbol{X}$ to provide estimates for the right-hand side of equation (4.4) and hence $\delta(\boldsymbol{x}_i)$. While the treatment effect of interest is expressed in terms of a difference in probabilities when the outcome is binary, this may not be the most appropriate scale to model treatment effect. The reason that the logistic model might provide a parsimonious model to predict the required conditional probabilities, is that a constant effect on the log odds scale has a valid interpretation across the entire range of predicted probabilities. For instance, the effect of treatment on outcome risk could really be constant on the log odds scale regardless of outcome risk in absence of treatment. This property does not hold for linear probability models or relative risk models, but very similar results may be obtained for probit models. As a quick reminder, Figure

4.1 shows the logistic link that transforms the log odds or linear predictor scale to the probability scale. A constant treatment effect on the log odds scale has a large effect on the probability scale when the linear predictor equals zero and approaches zero for very low and high linear predictor values. This nicely reflects the difference in the amount of wiggle room when the control outcome risk reflects unpredictability versus near certainty respectively. The reason to emphasize these well-known properties is to highlight that a very simple model on the log odds scale may very well lead to potentially relevant differences between individuals on the level of $\delta(\boldsymbol{x}_i)$ (*i.e.* differences in absolute risk).

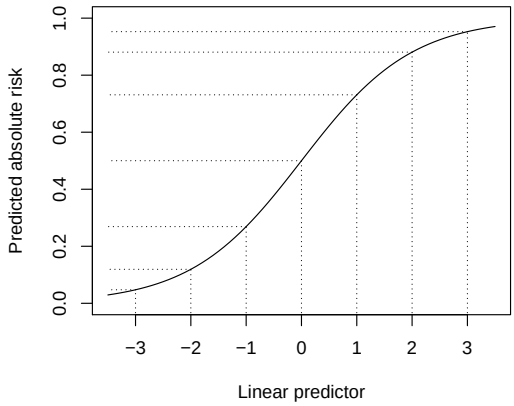


Figure 4.1: The translations of effects from the linear predictor (LP) scale to the absolute scale vary depending on location as shown for the logistic link function ($\frac{1}{1+e^{-\mathrm{LP}}}$). Therefore, a constant or homogeneous relative effect, here shown as a 1-point difference on the linear predictor (*i.e.* log odds) scale for ease of exposition, has different implications on the absolute risk scale. For example, for a patient with a control risk of 50% (LP=0), a treatment effect of -1 on the log odds scale reduces predicted absolute risk to 27% (LP=−1, resulting in an absolute risk reduction of 23%). For a patient with a control risk of 27% (LP=−1), the same treatment effect on the log odds scale leads to a predicted risk of 12% (LP=−2, resulting in an absolute risk reduction of 15%).

The simplest way to create such a model is to assume absence of any interaction between treatment and the other covariates. Thus, the (conditional) treatment effect is assumed to be *homogeneous* or constant on the log odds scale. In terms

of notation, let Y_i denote the independent dichotomous *observed* outcomes, assumed to have patient specific mean μ_i (*i.e.* $Y_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\mu_i)$). Also, with vector $\boldsymbol{x}_i$ denoting the p individual characteristics and a_i being the treatment indicator as before, this leads to the following simple logistic regression model

$$\text{logit}(P(Y_i = 1 | A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) = \beta_0 + \beta_t a_i + \boldsymbol{\beta}^\top \boldsymbol{x}_i \tag{4.5}$$

with regression parameters β_0, β_t, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. The key assumptions for this logistic model including a conditional *homogeneous* treatment effect β_t are i) appropriateness of the logistic link function, ii) linearity in the parameters, and iii) additivity of at least the treatment effect on the log odds scale (*i.e.* there are no treatment-covariate interactions on the log odds scale). The linearity assumption on the covariate contributions could easily be relaxed, allowing for global or local transformations of $\boldsymbol{x}_i$ such as polynomials and splines.

The predicted individualized treatment effect follows directly after estimation of the model parameters [1]:

$$\hat{\delta}(\boldsymbol{x_i}) = \frac{1}{1 + e^{-(\hat{\eta}_i + \hat{\beta}_t)}} - \frac{1}{1 + e^{-\hat{\eta}_i}} \tag{4.6}$$

where $\hat{\eta}_i = \hat{\beta}_0 + \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$. Strict additivity of the treatment effect on the log odds scale may provide a parsimonious model for the analysis and is easily translated to the more interpretable scale of $\delta(\boldsymbol{x_i})$ where this additivity does not hold. Note that in case of such a homogeneous treatment effect on the log odds scale, variability in $\boldsymbol{x}_i$ is the driving force behind any differentiation on the level of $\hat{\delta}(\boldsymbol{x_i})$. This variability in $\boldsymbol{x}_i$ corresponds to variability in prognosis across individuals under control treatment (*i.e.* variability in $Y_i^{a=0}$).

4.4.2 Heterogeneous or differential treatment effect

As an extension of homogeneous treatment effect models, the relative treatment effect can also be allowed to depend on the other covariates. We will refer to such non-additivity of the treatment effect on the relative scale as heterogeneity

[1] For our current goal and scope, inferring individualized treatment effect from predictions under the relevant potential outcomes is the end-point. However, it is interesting to note that this type of conditional potential outcome predictions can also serve as input for substitution estimators aiming for more marginal estimates, such as the parametric g-formula [22] and estimators of marginal risk difference and marginal risk ratio based on logistic models [100].

of treatment effect (HTE) or differential treatment effect. We note that there is no single accepted definition of the term HTE in the literature, and that it is sometimes used in a broader sense to also include the variability in $\delta(\boldsymbol{x_i})$ that may result from a homogeneous treatment effect model [89]. We use HTE in its narrow sense (*i.e.* restricted to non-additive relative effects), since this allows one to distinguish between possible variability in the way treatment affects different individuals (homogeneous versus heterogeneous) and variability amongst individuals that does not relate to treatment effect (*i.e.* variability in expected prognosis under control treatment $P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x_i})$).

The homogeneous treatment effect model in equation (4.5) can easily be extended to allow for HTE by inclusion of treatment-covariate interactions. The model then becomes

$$\text{logit}(P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) = \beta_0 + \beta_t a_i + \boldsymbol{\beta_m}^\top \boldsymbol{x}_i + \boldsymbol{\beta_z}^\top \boldsymbol{z}_i a_i \tag{4.7}$$

where $\boldsymbol{z}_i$ is a subset of $\boldsymbol{x}_i$, $\boldsymbol{\beta_m}$ includes the coefficients for the main effects of $\boldsymbol{x}_i$, and $\boldsymbol{\beta_z}$ includes the coefficients for treatment-covariate interactions. As before, the space of the measured $\boldsymbol{x}_i$ can be expanded using global or local transformations. These transformations need not be the same in $\boldsymbol{z}_i$: the functional form of the effect of a covariate may depend on treatment status. Also, note that when $\boldsymbol{z}_i$ equals $\boldsymbol{x}_i$ (*i.e.* all covariates are involved in treatment-covariate interactions), an exactly equivalent parametrization can be obtained by specification of separate models for the treated group and the control group with just an intercept and main covariate effects, which separates the models for both potential outcomes. Additional details are provided in the supplementary material (Part S4.1).

Based on the model in equation (4.7), the predicted individualized treatment effect again follows easily from the parameter estimates. In analogy to equation (4.6), the prediction of $\delta(\boldsymbol{x_i})$ can be derived according to

$$\hat{\delta}(\boldsymbol{x_i}) = \frac{1}{1 + e^{-(\hat{\eta}_i + \hat{\beta}_t + \hat{\boldsymbol{\beta}}_{\boldsymbol{z}}^\top \boldsymbol{z}_i)}} - \frac{1}{1 + e^{-\hat{\eta}_i}} \tag{4.8}$$

While this prediction comes nice and easy in theory, the challenge lies in precise model specification and the estimation of the model parameters of these relatively complex models.

4.5 Model estimation

The preferred method for estimation of the prediction models of interest depends on the relation between model complexity and the amount of signal in the data. In medical statistics, the amount of variability in the outcome of interest that can be explained is often low to moderate, which is the main reason for the need for large sample sizes. In case of insufficient sample size, models are prone to overfitting, which describes the situation where the model captures part of the noise in the data. Overfitting is an important concern since it limits generalizability of the model. In this section we discuss the need for methods that mitigate the susceptibility to overfitting.

4.5.1 Maximum likelihood

Estimates for the proposed logistic regression models can be derived with standard maximum likelihood estimation for generalized linear models [101, 102]. However, even after taking all content knowledge into account, models are often still overly complex with respect to the available sample size. This may already hold for a logistic model of homogeneous treatment effect when the number of covariates is large, or their functional form allowed to be complex and the sample size is relatively small [103]. It has long been recognized that standard maximum likelihood estimation of logistic models is problematic in these settings due to finite sample bias, perfect separation, collinearity, and overfitting [99, 104]. Models including many non-additive effects such as treatment-covariates interactions are even more prone and necessitate either strong prior assumptions or restrictions during model estimation.

4.5.2 Penalized maximum likelihood

Regression based on penalized maximum likelihood estimation (pML) has emerged as a method that can, at least to some degree, cope with relatively complex models [62, 64, 105]. These methods penalize the log-likelihood for the magnitude of regression coefficients other than the intercept. This penalty introduces bias towards zero on the estimated (non-intercept) coefficients, and thus towards the overall outcome incidence for predictions. In other words, it introduces a bias that reduces variability at the level of the predictions. This balance is also known as the bias-variance trade-off. Well-known penalized maximum likelihood methods include ridge regression, lasso regression, and the elastic net which includes the first two as special cases [65]. The ridge penalty is a smooth

penalty on squared size of the regression parameters and leads to shrinkage of the estimated coefficients. It was originally developed to deal with collinearity and tends to distribute the weight amongst collinear variables [62]. The lasso penalty is on the absolute value of the coefficients and leads to both shrinkage and selection [64]. It has a tendency to select amongst collinear variables. The elastic net penalty is a weighted balance of the ridge and lasso penalty. The required degree of penalization is a tuning parameter that needs to be estimated from the data, which is most commonly done by means of cross-validation. Importantly, this estimation involves uncertainty that is most problematic when accurate penalization is needed the most (*i.e.* small data sets and/or low signal relative to noise) [106, 107]. With respect to the equivalence of a model including all treatment-covariate interactions on the one hand and separate modeling in the treated and control group on the other hand (section 4.4.2), note that this equivalence no longer holds in case of penalized maximum likelihood. Details are provided in the supplementary material (Part S4.1).

Shrinkage and/or selection

Penalization in clinical prediction modeling
In some settings, the underlying process to be modelled is fairly well know, and therefore, the same holds for the elements that should be included in a model. While such a setting does not require selection of parameters, shrinkage may still be beneficial in terms of prediction accuracy when sample size is limited with respect to model complexity. In contrast, the available data may be very rich while the underlying process not well understood, but thought to be sparse. Penalization approaches that provide selection (*i.e.* sparse solutions) have been successfully applied to prediction problems with many covariates (possibly more than cases, *i.e.* $p >> n$) where selection of variables is key and there is insufficient content knowledge to do so, such as the selection of possibly important signals from microarray data [61].

In the typical context of clinical prediction modeling, the properties of the problem are somewhere in-between these two extremes. In the best-case scenario, clear pre-specification of a model might be possible. Often however, even though the data are typically low-dimensional, some further variable selection may still be required. The choice of penalty, and hence the need for selection, therefore heavily depends on the state of content knowledge and the amount of data available. An issue that may guide the selection of a penalty function is the need for an honest representation of the relative weights of model parameters.

Ridge regression tends to share the regression weights between parameters that are correlated with the outcome *and* with each other [63], while lasso tends to select amongst such variables [61, Chapter 4]. As an example, if two highly correlated covariates are equally predictive of the outcome, ridge will keep both in the model with approximately equal weight. Lasso will remove the one variable that happens to have a slightly weaker association with the outcome in the current sample. Both representations can be useful, but they serve a different purpose.

Penalization and modeling of heterogeneous treatment effect
Models of heterogeneous treatment effect include treatment-covariate interactions. Such interactions are always harder to estimate than the overall (main) treatment effect. In terms of selection, heterogeneous treatment effect models encounter the variable selection issue twice: for the main effects and for the treatment-covariate interactions. Usually, lower order (main) effects are kept in the model for each component of an interaction. However, both lasso regression and elastic net regression do not respect this hierarchical nature. To that effect, a hierarchical group lasso algorithm has been developed that does respect the hierarchy between main effects and interaction effects [108]. In short, variable selection is achieved on a group level that is allowed to be hierarchical, such that main effect groups will be in the model when they are part of any interaction. For problems with non-overlapping groups, regular group lasso algorithms are also widely available [109].

4.6 Model complexity

The current state of subject matter or content knowledge, the type of process under study, the strength of the associations in the data, and the final purpose of the model, all weigh into the decision on the best balance between prior model specification and more data-driven modeling methods. We have described both specification and estimation of logistic models that can be used to predict treatment effect. In this section, we will shortly touch upon the complementary roles of content knowledge and penalization, the concern of overfitting when aiming for out-of-sample prediction, and the degree to which model complexity can be left to the data-drive methods.

4.6.1 Content knowledge and penalization

Content knowledge, general statistical knowledge, and data-driven methods such as penalization should work synergistically to arrive at the most parsimonious model that reflects current content knowledge as updated by the data. Ideally, content knowledge includes knowledge on non-linear covariate contributions, interactions amongst covariates, and especially knowledge on covariates that might interact with treatment. Unfortunately, such knowledge is often very limited. Consequently, the number of parameters that one wants to estimate may still be relatively large even after all content knowledge is exhausted. While data driven approaches can help to nudge such models in the right direction, they provide no panacea or substitution of content knowledge.

4.6.2 Overfitting and prediction accuracy

For our aim to predict individualized treatment effect, the conditional model of treatment effect is intended to generalize and therefore overfitting is a key concern. The primary challenge is to balance model complexity with respect to the amount of available data in a way that generalizes beyond the original sample. The preferred way to do so is to limit model complexity based on content knowledge, supplemented by use of penalization. The relation between prediction accuracy, model complexity, effective sample size, and the strength of the associations in the data is well known and has been described in the context of risk prediction (*e.g.* [104, 103]). However, estimation of $\delta(\boldsymbol{x}_i)$ requires the difference between two outcome risk predictions (under the two to be compared treatments) to be accurate. While these two predictions will be highly correlated since they arise from the same individual, the expected error will invariably be larger than for a single prediction. To our knowledge, there is no guidance available on the necessary conditions with respect to effective sample size and expected explained variation for accurate prediction of risk differences within individuals. Our simulation study below provides some first insights.

4.6.3 'Risk modeling' versus 'effect modeling'

The use of treatment-covariate interactions to model heterogeneous treatment effect (*e.g.* as in equation (4.7)) has also been referred to as 'effect modeling' and has been distinguished from 'risk modeling' [89, 90, 110, 111, 112]. In risk modeling, treatment effect variability is evaluated as a function of outcome risk, where outcome risk is a function of the covariates. Therefore, it can be seen

as a data reduction method. For example, in risk modeling the linear predictor scores $\hat{\eta}_i$ resulting from a model for $P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x_i})$ can be interacted with treatment instead of the full multi-dimensional representation of $\boldsymbol{x_i}$. That is,

$$\text{logit}(P(Y_i = 1|A = a_i, \hat{\eta}_i) = \beta_t a_i + \hat{\eta}_i + f(\hat{\eta}_i)a_i \tag{4.9}$$

where $f(\cdot)$ denotes a possibly flexible function and $\hat{\eta}_i$ is used as an offset. The general idea is that $f(\hat{\eta}_i)$ is a much simpler structure to estimate than many individual treatment-covariate interactions. Therefore, it is supposed to fill a gap in situations where content knowledge is insufficient to limit model complexity to something that can be reliably estimated in the data.

From a statistical point of view, 'risk modeling' does of course reduce the risk of overfitting, since it restricts the modeled treatment effect heterogeneity to be a function of a scalar. However, the price to pay is that HTE is thereby forced to be proportional to the main effects of $\boldsymbol{x_i}$. This implies i) that all covariates that have a main effect also modify the relative treatment effect, and ii) that the effect of each element of $\boldsymbol{x_i}$ on HTE has the same direction as its effect on outcome risk. These are strong assumptions that have no clear biological substrate and no clear statistical preference over other data reduction methods such as principal component analysis [2]. Nonetheless, the idea behind risk modeling does reflect recognition of the danger of overfitting when modeling many treatment-covariate interactions, which remains an important issue when modeling heterogeneous treatment effect.

4.6.4 Tree-based methods

Many different modeling techniques are available under the machine learning umbrella. While we have limited our scope to regression-based methods, there are other methods that could be used instead. In particular, several tree-based

[2]The preference to model HTE as a function of outcome risk was originally motivated as "outcome risk is a mathematical determinant of treatment effect" [89], along with a reference to the fact that the [marginal] odds ratio [of exposure to treatment] equals $EER/(1-EER) \div CER/(1-CER)$ [where EER is the experimental outcome prevalence and CER is the control outcome prevalence]. However, the same equation can be used to show that independence is a possibility: i) only the $CER/(1-CER)$ part depends on control outcome risk and is positive and finite for any control prevalence in the $(0, 1)$ interval, and ii) combined with any valid odds ratio, $EER/(1-EER)$ is also positive and finite and therefore maps back to a prevalence amongst the treated in the $(0, 1)$ interval. This also holds for the conditional treatment effect and was in fact one of the reasons to prefer logistic regression as explained in section 4.4.1.

methods have been developed for the specific purpose of individualized treatment effect prediction. While an in-depth discussion is beyond our scope, we here provide several key references. Wager et al. [113] extend the well-known random forest algorithm by Breiman [114] to enable individualized treatment effect prediction and provide a very thorough overview of the required conditions for causal inference, including asymptotic theory. Also, they provide an overview of the literature on forest-based algorithms for estimating heterogeneous treatment effect. Lu et al. [115] provide a clear exposition of individualized treatment effect prediction and the potential outcome framework, and provide empirical and simulation results on a wide variety of random forest methods for causal inference, including virtual twins, counterfactual random forests, the aforementioned causal forests and Bayesian adaptive regression trees. Generalized random forests [116] constitute a more recent addition to the literature and form a much broader method that can also be used for causal individualized treatment effect estimation (and effectively encompass the work by Wager et al. [113]). Lastly, model-based recursive partitioning is a somewhat different tree-based approach that incorporates parametric models into the tree [117]. Such a parametric model can for instance describe control outcome risk and relative treatment effect. Node-splitting then occurs on the variable that generates most instability in the parameters for this model. Seibold et al. [118] have developed a model-based recursive partitioning random forest to identify treatment effect heterogeneity. Model-based recursive partitioning is most closely related to the methods discussed in this tutorial since it can differentiate between heterogeneity of treatment effect on the relative scale and differential outcome risk that may be related to a homogeneous treatment effect (i.e. constant log odds of treatment).

Beyond the predictions of individualized treatment effect, there have been efforts to find subgroups that might need different treatment based on a fitted causal forest [119], and efforts to further explain random forest-based treatment effect predictions based on covariate data [115]. Both papers go through considerable lengths to disentangle and interpret predicted individualized treatment effect differences. While this may lead to interesting hypotheses, this should be a careful undertaking. An illustration of the things that could go wrong is available in Rigdon et al. [120], showing high false discovery rates if such care is not taken.

To the best of our knowledge, direct comparisons of regression-based and other methods for the prediction of individualized treatment effect have not been performed yet. In general, it can be expected that more flexible models are more

prone to overfitting and require more data [121]. A challenge in the comparison of different methods in simulation studies is that specific data generating mechanisms favor specific methods. For instance, a data generating mechanism with linear and additive effects will favor regression methods; a mechanism generating subgroups based on cut-offs will favor tree-based algorithms. An interesting method to acknowledge these effects and check robustness is provided by Austin et al [122], and could also be applied in the context of individualized treatment effect prediction in future work comparing a wider set of methods.

4.7 Learning from simulations

We conducted a simulation study to illustrate the consequences of the choice of model and estimation method when predicting individualized treatment effect. Such simulations are especially helpful when predicting potential outcomes, since they can never be observed directly in practice. Also, the ability to manipulate the data generating mechanism into several interesting settings has a clear illustrative advantage. In the design of our simulation study, we adhered to the general guidelines proposed by Burton et al. [123] and Morris et al. [124].

4.7.1 Data generating mechanisms

The data generating mechanism was parametric and was based on a logistic model (equation (4.7)). Settings varied across the full factorial combination of varying total sample size (400, 1200, 3600), presence/absence of a main treatment effect ($\beta_t = ln(0.6)$ or $\beta_t = ln(1)$), and presence/absence of heterogeneity of treatment effect. The treatment indicators a_i were independent samples from a Bernoulli distribution with probability 0.5. Twelve covariates were drawn from a multivariate standard normal distribution with a compound symmetric covariance matrix ($\rho = 0.1$). Main effect coefficients $\boldsymbol{\beta_m}$ were of exponentially decreasing size ($\beta_{m,1}, \ldots, \beta_{m,12} = 2^{-\frac{0}{2}}, 2^{-\frac{1}{2}}, 2^{-\frac{2}{2}}, \ldots, 2^{-\frac{11}{2}}$) to reflect i) decreasing added value of consecutive variables, and ii) that it is unlikely for variables that are included in a risk model to have truly zero coefficients. In settings with a homogeneous treatment effect, there were no treatment-covariate interactions (*i.e.* $\boldsymbol{\beta_z} = \mathbf{0}$). In settings with a heterogeneous treatment effect, $\boldsymbol{\beta_z}$ was equal to $-\frac{1}{2}, -\frac{1}{4}$ and $-\frac{1}{8}$ for $\beta_{z,10}, \beta_{z,11}$ and $\beta_{z,12}$ respectively, and included a small random perturbation that was generated once for all simulations ($\leq |0.05|$) for $\beta_{z,1}, \ldots, \beta_{z,9}$. In each simulation setting, the intercept was chosen such that the true underlying outcome prevalence in the control arm was 25% (details

provided in the supplementary material, Part S4.2). Nagelkerke R^2, as measured between the true conditional probabilities for the assigned treatment (*i.e.* $P(Y_i^{A=a_i} = 1|\boldsymbol{X} = \boldsymbol{x}_i)$ and the observed events Y_i in a large sample, was approximately 0.4 for all settings.

To get more insight into the different simulation settings, the distribution of $\delta(\boldsymbol{x}_i)$ for each of the data generating mechanisms is shown in Figure 4.2. Note that i) there is no variability at all in the upper right figure (due to absence of any treatment effect); ii) variability in the upper left figure is due to variability in control risk across patients, and iii) variability in the lower two figures is due to heterogeneity in both control risk and treatment effect.

4.7.2 Model development

Within each simulation run, both a development and a validation data set were simulated for each of the simulation settings. That is, both data sets were always generated according to the same data generating mechanism. The size of the model development sets matched the simulation settings (*i.e.* 400, 1200 or 3600), and the size of the validation sets was always equal to 10,000 observations.

Table 4.1 provides an overview of the evaluated methods. The overall absolute treatment effect is just the marginal version of $\delta(\boldsymbol{x}_i)$ as marginalized over all covariates. Its estimate of $\delta(\boldsymbol{x}_i)$ is just the $\hat{\delta}$, the difference in mean outcome incidence between treatment arms. For the homogeneous treatment effect models, all covariates entered the model only as main effects. In case of heterogeneous treatment effect models, all covariates entered the model as both main effects and interactions with treatment. The selected penalty parameter for ridge, lasso, and hierarchical group lasso (HGL) was the one with the smallest deviance in 10-fold cross-validation. Note that both lasso and HGL may set coefficients to exactly zero (*i.e.* perform selection). Also, while HGL can search the entire interaction space, the current implementation was limited to treatment-covariate interactions in parallel to the other methods. A final variation of the HTE methods was a 'content knowledge' (CK) setting, where we assumed that content knowledge suggests that only the first eight variables have important main effects, and only covariates nine to twelve are likely treatment interaction candidates. Ridge regression was used to estimate such a CK-based model. The 'risk modeling' implementation included a risk model estimated in the control group based on main effects for all covariates and a linear treatment with risk-score interaction. Both standard maximum likelihood and ridge regression were performed for the risk model. A significance-based approach

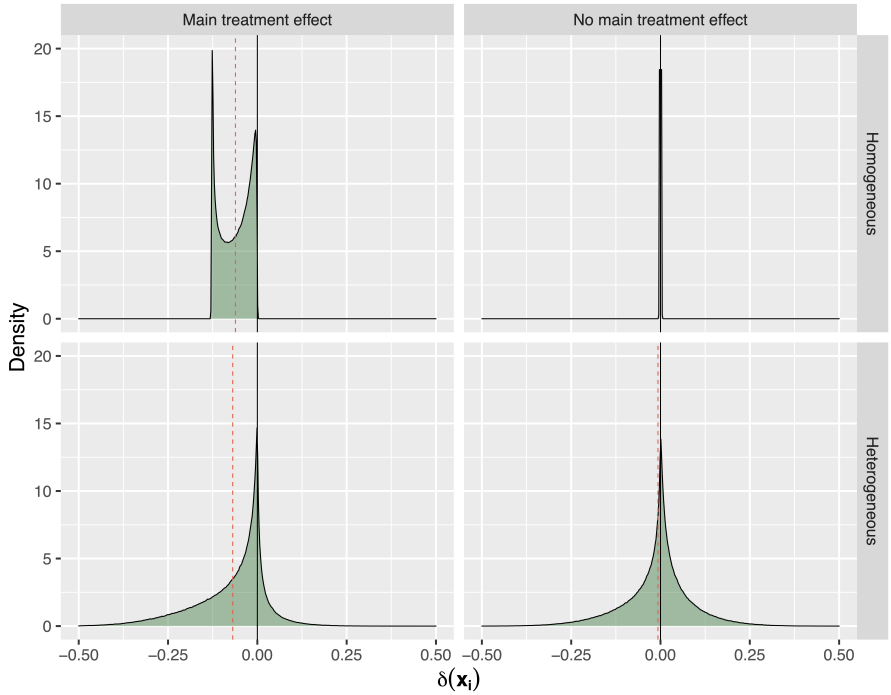


Figure 4.2: Distribution of $\delta(\boldsymbol{x}_i)$ according to the data generating mechanism for each of the simulation settings. The quadrants correspond to settings with a homogeneous main treatment effect (upper left), absence of any treatment effect (upper right), heterogeneous treatment effect in presence of a main treatment effect (lower left), and heterogeneous treatment effect in absence of a main treatment effect (lower right). The red dotted lines provide the mean of $\delta(\boldsymbol{x}_i)$ per setting. Note that all of the mass in the upper right figure is on a spike at $\delta(\boldsymbol{x}_i) = 0$.

was implemented as a final comparison. Starting from a homogeneous treatment effect model including all covariates, a likelihood ratio test was performed for the treatment effect coefficient. When non-significant, the treatment coefficient was removed from the model (leaving only main covariate effects). When significant, all treatment-covariate interactions were added to the model and a second likelihood ratio test was performed to evaluate their joint significance.

Treatment-covariate interactions were kept in the model when this joint test was significant and removed otherwise. All tests used an α level of 0.05. In addition to the methods in Table 4.1, HTE was modeled using separate prediction models per treatment arm, and thus per potential outcome, as described in the supplementary material (Part S4.1).

Regression model	**Equations**	**Estimation method**
Overall absolute treatment effect (Overall)	–	ML
Homogeneous treatment effect (HOM)	(4.5) (4.6)	ML, ridge
Heterogeneous treatment effect (HTE, HTE-CK)*	(4.7) (4.8)	ML, ridge†, lasso, HGL
'Risk modeling' (RM)	(4.9)	ML, ridge
Significance-based (SB)	–	ML

Table 4.1: Implemented methods towards the prediction of individualized treatment effect ($\hat{\delta}(\boldsymbol{x}_i)$).
* Model specification differs between the default case (HTE) modeling all main effects and treatment-covariate interactions, and the content knowledge case (HTE-CK) modeling a selection of main effects and an treatment-covariate effects (details are described in Section 4.7.2).
† Only ridge was used for the HTE-CK model.
Abbreviations: ML (Maximum Likelihood), HGL (Hierarchical group lasso).

4.7.3 Model evaluation

Each of the methods provides a prediction vector $\hat{\boldsymbol{\delta}}$ with elements $\hat{\delta}_i$ (short for $\hat{\delta}(\boldsymbol{x}_i)$) as derived in the validation sample. These predictions can be compared to the known $\boldsymbol{\delta}$ based on the data generating mechanism. The root mean squared prediction error (rMSPE) between the elements of these two vectors was used to quantify the prediction errors according to

$$\text{rMSPE} = \sqrt{n^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})^\top(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})}. \tag{4.10}$$

The root was taken to arrive at an expression of the error on the risk difference scale. In addition, the 0.9-quantile of absolute prediction errors was derived

for both $\hat{\boldsymbol{\delta}}$ and predicted risk. Supplementing these single figure summaries, calibration plots were derived for both $\hat{\boldsymbol{\delta}}$ and predicted risk, where predictions were cut into twenty equal-size quantiles groups and compared to the true values based on the data-generating mechanism.

4.7.4 Statistical software

All analyses were performed in `R` statistical software version 3.5.0 [125]. The `R` script for exact replication of the simulation study is available as online supplementary material. Logistic regression models based on maximum likelihood estimation were fitted using `glm()`. Ridge and lasso implementations were based on the `glmnet` package [65]. Hierarchical group lasso was implemented using the `glinternet` package [108].

4.7.5 Simulation study results

We here synthesize the results of 250 simulation runs in terms of root mean squared prediction error and calibration of the predicted individualized treatment effect.

Average root mean squared prediction error

The main simulation results with respect to root mean squared prediction error (rMSPE) are shown in Figure 4.3, which provides a summary of the error that can be expected in the long run across settings and methods. Five key observations can be made across all settings. First, conditioning on main effects of the available covariates as in a homogeneous treatment effect model was always beneficial when compared to the fully marginal $\hat{\delta}$. This confirms the idea that a conditional estimand ($\delta(\boldsymbol{x}_i)$) requires a conditional estimator. Second, HTE model accuracy was especially sensitive to sample size, which is in line with expectation due to the amount of parameters that needs to be estimated. Third, penalization is key and improved estimation of both homogeneous and heterogeneous treatment effect models up to large sample sizes. Fourth, penalization does *not* remove the risk of overfitting for complex models in small sample size settings. Heterogeneous treatment effect models could not be reliably fitted in small samples regardless of the estimation method. Fifth, utilization of content knowledge was the most effective way to reduce HTE model complexity. Lasso and HGL did not catch up, even though the content knowledge simulated here was not entirely correct and the lasso models did start

from the correct set of variables given the data generating mechanism. The reduction in the potential set of treatment-covariate interaction variables in the content knowledge-based model was very effective, even though this missed out on small interaction effects. Nonetheless, apparent content knowledge could of course not help out when it was wrong, such as in the complete null setting (no main treatment effect, no heterogeneity). Lastly, risk modeling was not the preferred method in any of the simulation settings. This of course depends on the data generating mechanism where HTE was not fully explained by a risk model, but this would also not be expected in practice and serves the purpose of showing that risk modeling may miss important treatment effect heterogeneity. Additional rMSPE results for comparing treatment-covariate interaction modeling with separate prediction models per treatment arm, and thus per potential outcome, are described in the supplementary material (Part S4.1). In short, treatment-covariate interaction modeling by means of lasso regression was best across all settings when compared to per arm modeling. The differences between treatment-covariate interaction modeling and per arm modeling were more nuanced for ridge regression.

Calibration

The online supplementary material (Part S4.3) provides calibration plots of $\hat{\delta}(\boldsymbol{x}_i)$ across methods and settings. It provides further insight into the distribution of the errors and the degree of variability across replications. Several important observations can be made that supplement the conclusions based on rMSPE. First, calibration curves at least pass the $(0, 0)$ point, and the size of the errors increases as predictions move away from zero. Therefore, prediction errors were much smaller around the harm-benefit boundary (*i.e.* $\delta(\boldsymbol{x}_i) = 0$). Second, calibration in individual small samples could be far off even if average performance across simulations was good. For instance, fitting a ridge regression homogeneous treatment effect model in accordance with the data generating mechanism could still lead to substantial overfitting or underfitting in any individual data set. These findings on the risk difference scale are in line with earlier results on direct prediction [106, 107]. Third, in small and even medium sample sizes, even penalized heterogeneous treatment effect models overfitted to such an extent that they falsely predicted harm for a subpopulation when in fact there was none (as in the homogeneous treatment effect settings). Fourth, when the data generating mechanism was heterogeneous, calibration of predicted treatment effect was quite reasonable for penalized HTE models in medium and large sample sizes. Note that, while useful for illustrative purposes in this simulation

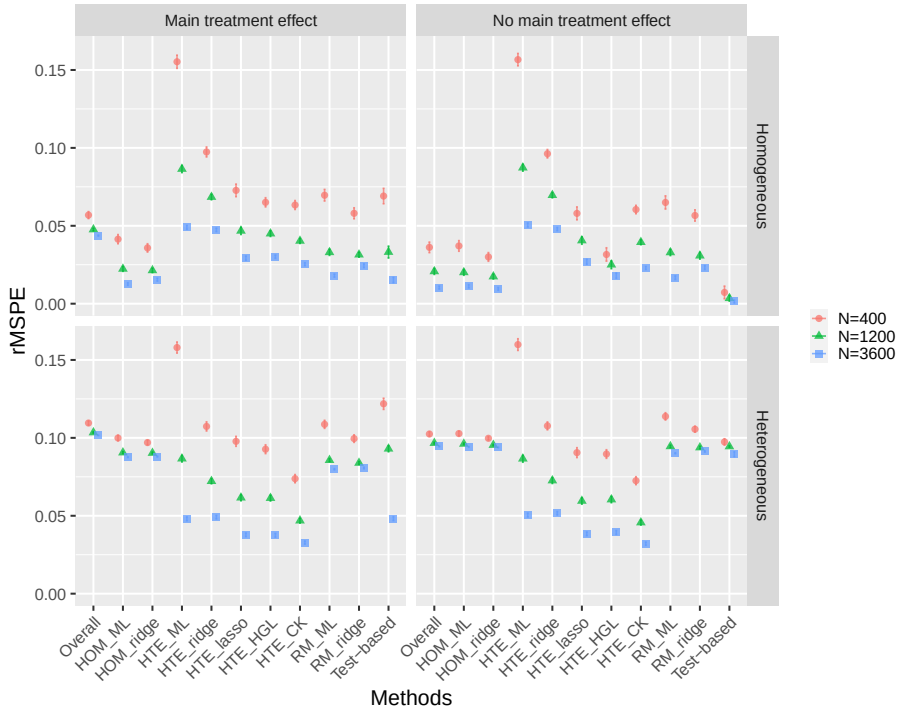


Figure 4.3: Simulation study results: average root mean squared prediction error (rMSPE) of the predicted treatment effects (over 250 simulations) with ±2 SE error bars for all simulation settings. The quadrants correspond to settings with a homogeneous main treatment effect (upper left), absence of any treatment effect (upper right), heterogeneous treatment effect in presence of a main treatment effect (lower left), and heterogeneous treatment effect in absence of a main treatment effect (lower right). Note that the standard errors are often so small that they are obscured by the mean estimates. Abbreviations: homogeneous treatment effect models (HOM), heterogeneous treatment effect models with treatment-covariate interactions (HTE), and heterogeneous treatment effect models based on risk modeling (RM), as estimated by means of maximum likelihood (ML), ridge or lasso regression, or hierarchical group lasso (HGL), with CK standing for content knowledge.

setting, these calibration plots are not available in practice since they compare predictions against the true individual treatment benefits (which do not even have an observed individual level equivalent).

4.8 Applied examples

4.8.1 Acute otitis media

For illustrative purposes, we analyzed data from a randomized, double-blind, placebo-controlled trial of amoxicillin for clinically diagnosed acute otitis media (AOM) in children 6 months to 5 years of age [126]. This trial included 512 children and collected baseline data on antibiotic treatment received, sex, presence of recurrent AOM, fever, bilateral occurrence, ear pain, presence of a runny nose, cough, tympanic membrane abnormality, and age. All variables but the latter were dichotomous. The endpoint analyzed here is the same as reported by Rovers et al. [127]: positive when either fever or ear pain was present after 3 days of follow-up. While not truly binary, composite endpoints occur frequently in practice. Thus, data were available on a total of 9 patient characteristics, treatment, and on a composite dichotomous endpoint. All in all, there were 147 events.

We first fitted a logistic regression model on the full data set with main effects for treatment and the 9 patient characteristics. The estimated log odds of treatment was statistically insignificant, but in the expected direction ($\hat{\beta}_t = -0.34$, se=0.20). The apparent Nagelkerke R^2 for this model was only 8.8%, and a larger sample size would be generally be required for prediction model development in such low signal settings [12]. Considering the simulation study results, the sample size, and the low amount of signal with respect to predicted risk, the starting point for any prediction of individualized treatment effect in these data is very weak. Nonetheless, it is interesting to see whether the proposed methods indeed show a lack of predictive ability. To that effect, we internally validated all of the methods evaluated in the simulation study (except for the use of content knowledge) in 100 bootstrap samples. The lowest out-of-sample Brier score (0.202) was obtained with the homogeneous treatment effect model fitted by means of ridge regression. However, the accompanying out-of-sample Nagelkerke R^2 was near-zero and even negative for more flexible models. Together, these findings show a lack of strong support for a non-zero average treatment effect and for any ability to personalize treatment effect.

4.8.2 International Stroke Trial

The International Stroke Trial (IST) was a large randomized open trial comparing no antithrombotic treatment, aspirin treatment, and subcutaneous heparin treatment in a total of 19,435 patients with acute ischaemic stroke [128]. The individual patient data from the IST trial are available for public use and can be downloaded from the web [129]. For the current applied example, we used data from patients randomized between no treatment ($n = 4,860$) and aspirin treatment ($n = 4,858$). Interestingly, the effect of aspirin treatment on the combined endpoint of death or dependency at 6 months was evaluated conditional on a prognostic score in the original article (and found to be effective on average). The prognostic score was based on age, sex, state of consciousness, and 8 other neurological symptoms evaluated at baseline. All variables except for age (continuous) and sex (dichotomous) are categorical variables with three levels. The total number of events was 6,043. Ninety-nine patients with a missing outcome were omitted; covariates were complete.

We first fitted a logistic regression model on the full data set with a main effect for treatment and main effects for the sex, state of consciousness and the eight neurological signs. Categorical variables were dummy coded. The estimated log odds of treatment in this model was -0.11 (se 0.048), corresponding to an odds ratio of 0.90, and the apparent C-statistic and Nagelkerke R^2 were 0.79 and 0.31 respectively. Even though the relative treatment effect is small, the presence of an average effect, the ability to explain a substantial part of the risk of an event, and the amount of available data provide a good starting position when aiming to individualize treatment effect prediction. In terms of the simulation study results, the expectancy is that a HTE model would pick up heterogeneity if it is present and that a homogeneous model would be preferred otherwise.

For illustrative purposes, we therefore examine the predictions from a ridge homogeneous treatment effect model (HOM-ridge) and a hierarchical group lasso HTE (HGL-HTE) model as fitted in the entire sample. In both cases, the penalty parameter with the lowest 10-fold cross-validation deviance was selected. Figure 4.4 shows the very high correlation (0.996) between risks predicted from either model (upper left panel). Nonetheless, the distribution of $\hat{\delta}(\boldsymbol{x}_i)$ is quite different for both models (upper right and lower left panel). The HOM-ridge model hardly discriminates w.r.t. treatment effect, whereas the HTE-HGL model predicts harm for a substantial part of the population. The lower right panel shows the relation between $\hat{\delta}(\boldsymbol{x}_i)$ as predicted from both models.

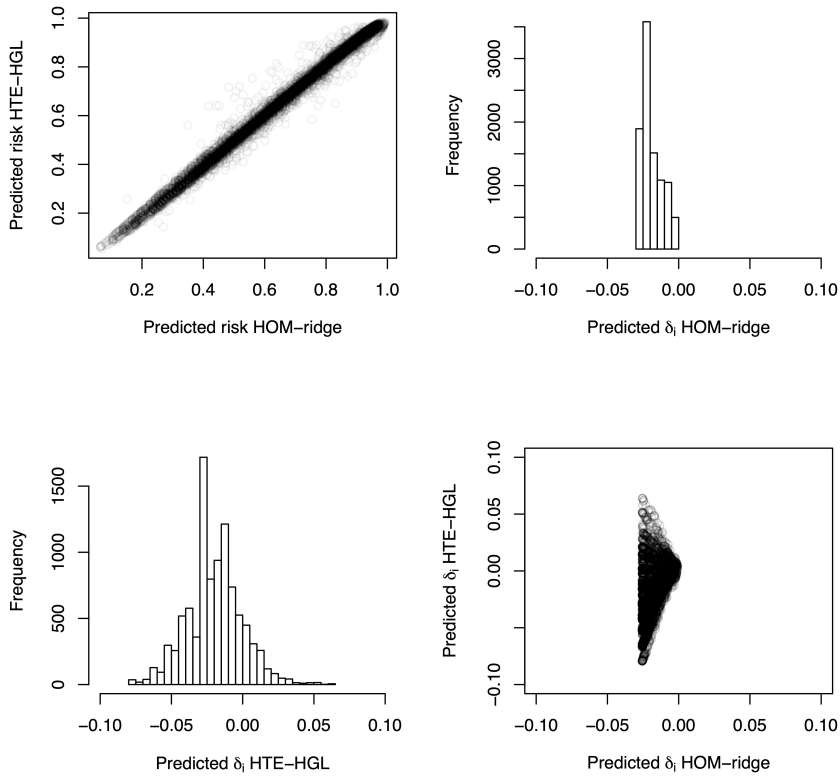


Figure 4.4: International Stroke Trial (IST) data. The upper left panel shows the relation between predicted risks based on the ridge homogeneous treatment effect model (HOM-ridge) and the hierarchical group lasso HTE (HTE-HGL). The histograms show the distribution of predicted individualized treatment effect for the same models. The lower right panel shows their mutual relation.

The main question is whether any, and if so, which of the predictions of $\delta(\boldsymbol{x}_i)$ can be expected to generalize. A key limitation is that $\hat{\delta}(\boldsymbol{x}_i)$ cannot be validated directly in real data. A possible approximation is to check whether groups based on quantiles of $\hat{\delta}(\boldsymbol{x}_i)$ relate to observed treatment effect within

these same groups. This is essentially a group level effort to approximate calibration at the level of $\hat{\delta}(\boldsymbol{x}_i)$. The idea is to make use of the fact that the potential outcomes, and thus $\delta(\boldsymbol{x}_i)$, are independent of treatment assignment (exchangeability). Figure 4.5 shows apparent and bootstrap results. As can be seen, the apparent calibration of predicted $\hat{\delta}(\boldsymbol{x}_i)$ is not good for HOM-ridge and reasonable for HTE-HGL. However, in both cases, bootstrap results show such a high degree of variation and lack of trend that the prediction of $\hat{\delta}(\boldsymbol{x}_i)$ cannot be trusted. In case of the HOM-ridge model, this may relate to the fact that such small variations in $\hat{\delta}(\boldsymbol{x}_i)$ cannot just not be retrieved from limited out-of-sample cases. In case of the HTE-HGL model, it implies that even in such a large sample, penalized models may still overfit.

To further illustrate this, we again applied all methods evaluated in the main simulation study (except for the application of content knowledge). We evaluated model fit by means of out-of-sample Brier score and Nagelkerke R^2 in 100 bootstrap replications. The best Brier score was achieved for the homogeneous treatment effect model fitted with standard maximum likelihood, closely followed by the equivalent ridge version and the hierarchical group lasso (HGL) and ridge HTE model (0.17910, 0.17915, 0.17921, and 0.17933 respectively). The same group of methods had the highest Nagelkerke R^2 values (0.3055, 0.3059, 0.3062, and 0.3066 respectively), with little difference between them. All in all, the differences in bootstrap-corrected estimates of overall fit were very small even though the number of model parameters differed substantially across models. Such results, and the extremely high correlation between risk predictions from different models (as in the left upper panel of Figure 4.4), may already be interpreted as red flags with respect to overfitting of the HTE models.

4.9 Practical considerations

In general, as with regular prediction modeling, overfitting is an important concern and sample size and penalization are key. We refer to recent guidance papers on sample size [104, 12, 103] and the use of penalization in clinical prediction modeling [106, 107]. These guidelines provide a lower bound for the sample size when developing models to predict individualized treatment effects. The required sample size will further increase when including treatment-covariate interactions.

With respect to the choice of modeling approach, the hierarchical group lasso (HGL) performed well across settings. HGL best captured treatment effect het-

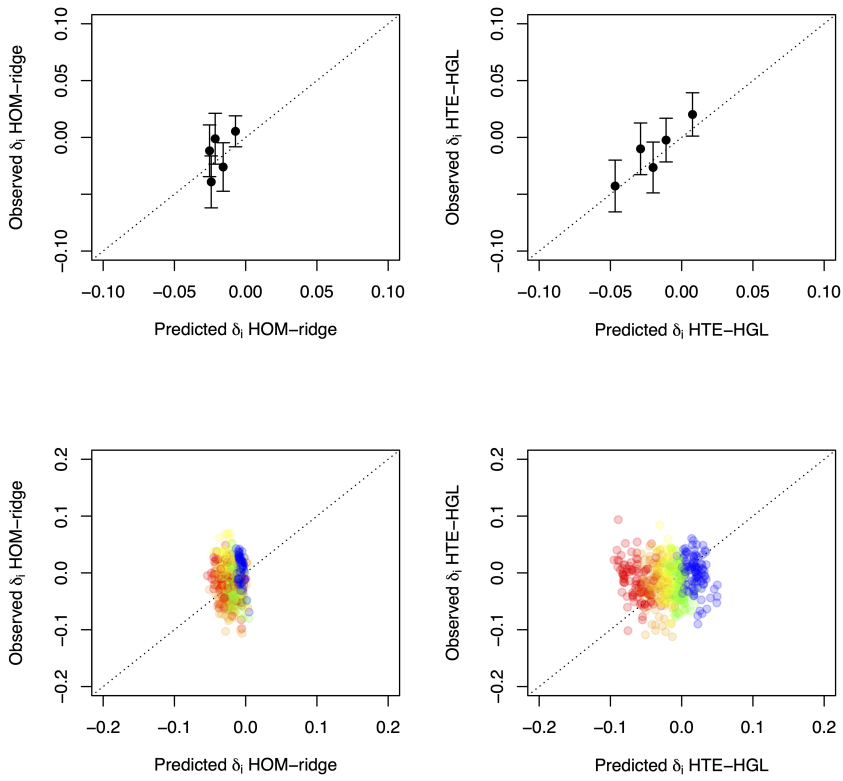


Figure 4.5: International Stroke Trial (IST) data. The upper panels show apparent calibration of treatment effect for the ridge homogeneous treatment effect model (HOM-ridge) and the hierarchical group lasso HTE (HTE-HGL). Dots describe mean $\hat{\delta}(\boldsymbol{x}_i)$ within quintiles of $\hat{\delta}(\boldsymbol{x}_i)$ (x-axis) versus the marginal treatment effect within those same groups (y-axis). Bars provide ± 1 SE. The lower panels show the out-of-sample estimates obtained from 100 bootstrap replications with red representing the lowest quintile $\hat{\delta}(\boldsymbol{x}_i)$, followed by orange, yellow, green, and blue for the highest quintile of $\hat{\delta}(\boldsymbol{x}_i)$.

erogeneity when present and had acceptable overfitting otherwise. It clearly outperformed HTE models based on ridge, lasso, or unpenalized maximum likelihood. In practice, a penalized homogeneous treatment effect model provides an important alternative model that is less prone to overfitting, but also less capable of capturing HTE. This modelling approach therefore appears particularly promising when RCTs are relatively small, or when there is (almost) no treatment-effect modification. Cross-validation or bootstrapping may be used to choose between HGL and homogeneous treatment effect models. The International Stroke Trial applied example provided an example of such a bootstrapping approach, showing the evaluation of well-known measures of overall fit (Brier score and Nagelkerke R^2) [51, 8], and a proposal to visually evaluate performance on the level of aggregated individualized treatment effect.

Nonetheless, samples that are very small with respect to the number of model parameters of interest should raise caution [104, 12, 103]. While penalization is clearly beneficial on average, accurate estimation of the penalization parameter itself cannot be expected in small sample size settings and there is no way to know whether this affects any particular model in practice [106, 107]. Also, the empirical evaluation of individualized treatment effect models is still in its infancy. Both group level evaluation of predicted individualized treatment effects and individual level evaluation of outcome risk predictions are only indirect approximations of the performance of individualized treatment effect prediction. Due to the insensitivity of the available measures, we expect model comparisons to be conservative in the sense that they will only prefer a heterogeneous model if the evidence is quite strong. For instance, HTE needs to be large enough to substantially affect overall outcome predictions, or sample size needs to be large enough to reliably assess aggregate predicted versus observed treatment effect. Also, complex models suffer more from the decreased variability in bootstrap samples.

Importantly, any measure based on observed data can only be an approximation of the performance of a potential outcome prediction model, with the remainder resting on assumptions [130]. The plausibility of the identifiability assumptions is an important part of model evaluation, including thoughts about their transportability, and requires thinking instead of measuring. To the knowledge of the authors, there is no specific guidance on the validation of the potential outcome type of individualized treatment effect prediction models beyond the guidance provided in this tutorial. This remains an important open area of research that needs further study.

Lastly, a note on the interpretation of any identified treatment effect heterogeneity. Randomization of treatment (or unconfoundedness after appropriate modeling of confounders in observational settings) supports causal inference with respect to treatment in a subgroup (some covariate status). This is subtly, but importantly, different from the assumption that subgroup membership causes or explains a change in treatment effect. For instance, subgroup membership might just be correlated with an unobserved underlying cause of treatment effect heterogeneity. Such implicit inversing of the causal relation underlies the exploration of subgroups as identified by individualized treatment effect prediction to i) inform treatment decisions, or ii) 'explain' treatment effect heterogeneity. However, such conclusions require randomization or unconfoundedness of subgroup membership. Care should be taken to emphasize that the estimated treatment effect describes the causal effect of treatment within a given subgroup, and not necessarily the other way around (*i.e.* it does not warrant the interpretation that the characteristics defining a subgroup cause differential treatment effect). Therefore, while predicted treatment effect heterogeneity may provide interesting hypotheses about its causal structure, it does not provide answers without further thinking about, and analysis of, the causal pathways involved.

4.10 Discussion

We have provided an overview of the process of individualized treatment effect prediction in the context of a randomized trial with a binary endpoint. To that effect, we have described the integration of key elements from the fields of causal inference and clinical prediction research. These methods can be used to expand on the mainstay analysis of randomized trials, and may help to uncover between-subject heterogeneity in terms of predicted outcome risk and treatment effect.

With respect to causal inference, we focused on the causal nature of the question of interest and a clear definition of individualized treatment effect based on the potential outcomes framework. From there, we explained the necessary assumptions to identify the individualized treatment effect based on the observed data. While such effects can in principle be estimated nonparametrically, further modeling is beneficial and allows straightforward comparison of treatment effect conditional on many covariates. Even though the prediction problem itself could be solved without any reference to causal inference, going through the motions increases clarity of the research question and gains understanding

of the requirements for a causal interpretation of the final model.

With respect to prediction modeling, we focused on the need for a parsimonious model with validity across the risk scale (log odds), while maintaining an interpretable scale for the final result (the risk difference scale). Specification of models for a homogeneous treatment effect (constant relative effect) and differential or heterogeneous treatment effect were described in detail. Subsequently, the relation between prior knowledge and data-driven methodology was examined, revealing the need for both. In line with general sample size guidance when developing a multivariable prediction model [12], sufficient sample size was important for accurate individualized treatment effects predictions and model stability.

While all of the required ingredients for individualized treatment effect prediction are well-known, their successful combination constitutes a challenging problem that is on the boundary of what can be observed in empirical data. Our simulation study, with a known data generating mechanism, provided clear insight into methods that are able to pick up heterogeneity in sufficiently large samples, while limiting the amount of overfitting in absence of heterogeneity. While very informative, actual analysis of observed data will have to rely on dichotomous outcomes from subjects observed under a single treatment condition that are only incompletely matched across treatment groups. The best way to evaluate the performance of individualized treatment effect prediction models is an open question. We described a bootstrap-based internal validation approach that decreases the risk of overfitting. A very recent contribution to the literature on potential outcome prediction and individualized treatment effect describes a very similar split sample approach [131]. Also, a novel type of c-index has been suggested to measure discriminative performance of individualized treatment effect predictions [132]. Nguyen et al. provide some cautionary notes on its interpretation and estimated standard error in their appendix [131].

The prediction of individualized or personalized treatment effect is an active field of research. Recent broad overviews on predictive approaches towards heterogeneity of treatment effect are available elsewhere and include a comprehensive overview of applied papers [133, 134]. Related work approaches the problem from the missing data perspective [135, 17]. Also, work has been done on individualized treatment effect prediction for optimal treatment selection [136, 137] and selection of patients for future studies [138, 139]. All in all, the literature on personalized medicine approaches that use prediction modeling is vast and too extensive to cite here. What we add is a clear, principled and from the

ground-up overview that integrates prediction modeling with causal inference and accentuates the importance of study design features.

To limit the scope, we did not venture into the incorporation of post-randomization measurements, dropout and selection bias, and observational data. Such topics require careful attention to the exchangeability assumption, which is no longer fulfilled by the study design and needs further assumptions and careful modeling with respect to all possible confounders. A recent scoping review provides an overview of the literature with respect to methods for causal prediction that extend to observational data [24]. Also, where we have focused on intention to treat estimates of point exposure treatment, different settings and questions require further thought on the relevant definition of the estimand [140]. As a second limitation, we provided a small simulation study covering a limited number of settings that was designed for illustrative purposes. The setup was such that development and test sets were generated from the same data generating mechanism. In practice, there may be differences between these two settings that are not captured by the models, and the uncertainty that accompanies these unknowns may overshadow relatively small gains realized by more complex models [141].

More general limitations pertain to the typical randomized trial design that provides the data to be used for individualized treatment effect prediction. Other designs, such as $N \times 1$ trials and cross-over designs may provide more direct within-person comparability, and thereby also provide information on the stability of treatment response for an individual [142]. However, these designs are infeasible for many conditions and have their own set of challenges [143]. More importantly, randomized trials are typically designed to be of sufficient sample size to reveal an anticipated average relative treatment effect. Therefore, randomized trials are not designed for complex prediction modeling. Hence, if we want to walk down the avenue of individualized treatment effect modelling, we will either have to design trials with this purpose in mind, or have to find more creative ways to amplify our data. This could include the analysis of individual patient data from multiple randomized trials, or even the use of non-randomized studies for the estimation of outcome risk under a control condition [24]. Besides clear opportunities, such approaches also bring about many new challenges. For instance, typical challenges that occur in clustered data settings (*e.g.* between-study heterogeneity) have been comprehensively illustrated in a recent tutorial on the examination of heterogeneous (relative) treatment effect in patient-level data from multiple randomized trials [144]. The implications of such challenges in the context of causal prediction research require further study.

4.11 Concluding remarks

We hope that our overview on the basics underlying individualized treatment effect prediction in binary endpoint settings is a useful guide and starting point for statisticians interested in this area. The successful implementation of individualized treatment effect prediction requires careful thought on the exact nature of the question and estimand(s) of interest, the causal and modeling assumptions relied on, and the ever-present bias-variance trade-off that requires even greater care than usual when working with potential outcomes. Sample size considerations are important in all areas of research and there is increasing awareness on the need for larger sample sizes when developing prediction models and examining treatment-covariates interaction. Future work is needed on the validation of models for predicted individualized treatment effect, their role in uncovering sources of heterogeneity, and ways to account for the clustered nature of many data sets. Also, beyond the frequentist framework, the basis for a fully Bayesian approach has long been recognized [145] and could combine the advantage of penalization with a more thorough view on the posterior distribution of the model parameters. A summary of key recommendations and findings is provided in Box 2.

Box 2: Key points and recommendations

- It is important to clearly define the individualized treatment effect of interest and to be aware of the identifiability assumptions underlying its causal interpretation.
- Analogous to causal inference with respect to average treatment effect, randomization of treatment greatly facilitates causal inference with respect to predicted individualized treatment effects.
- Logistic regression provides a parsimonious model to predict absolute individualized treatment effect (*i.e.*, treatment effect on the risk difference scale) in new patients. Even in absence of treatment-covariate interaction (*i.e.*, homogeneous patient-level treatment effect on odds ratio scale), a logistic model accounting for individual patient characteristics (prognostic factors) can lead to meaningful differentiation in terms of absolute treatment effect.

(continues on the next page)

Box 3: Key points and recommendations (Cont.)

- Sample size and penalized estimation are of key importance for accurate individualized treatment effect prediction; penalization alone does not guarantee accurate predictions in new individuals when sample size is insufficient with respect to model complexity. Existing guidelines on clinical prediction modeling provide a lower bound for the sample size needed in case of individualized treatment effect prediction [12, 104, 103].
- In practice, bootstrap internal validation of likelihood-based measures of overall fit (*e.g.* R^2, AIC), mean squared prediction error (*e.g.* Brier score), and aggregate observed versus expected measures of treatment effect variability (as in Section 4.8.2) help to choose amongst competing models. It is recommended to include a homogeneous treatment effect model as a starting point (reference model) and to consider more complex models based on biological, clinical and statistical evidence.
- Future work is needed to further delineate best practices for the evaluation of individualized treatment effect predictions and models, hence also improving model comparison and validation procedures.
- Purely explorative indications of heterogeneous treatment effect provide an interesting starting point for further research (e.g. into the causal structure of the heterogeneity) and require external validation.
- This tutorial handles causal prediction of treatment effect on a binary outcome, conditional on individual level patient characteristics. This should not be confused with a causal interpretation of the effect of individual level patient characteristics on the effect of treatment.

Acknowledgements

J Hoogland, M Belias, J in 't Hout, MM Rovers, TPA Debray and JB Reitsma acknowledge financial support from the Netherlands Organisation for Health Research and Development (grant 91215058). TPA Debray also acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050). We want to thank the researchers involved in the original otitis media [126] and stroke trials [128] for use of their data.

Data Availability

Data for the International Stroke Trial applied example are publicly available [129]. Data for the otitis media applied example not available for sharing since they contain privacy sensitive data according to the General Data Protection Regulation. `R` scripts to perform the simulation study, including data generation and analysis, are available for sharing.

Supplementary Material

The supplementary material consists of further information on separate prediction modeling per potential outcome (Part S4.1), the generation of data for a given outcome prevalence (Part S4.2), and calibration results for the simulation study (Part S4.3).

S4.1 Separate modeling of each potential outcome

This supplementary material describes the equivalence between a special case of the heterogeneous treatment effect model and models fitted separately in each arm of the trial (section S4.1.1), the loss of this equivalence when introducing penalization (section S4.1.2), and simulation results comparing treatment-interaction models with models fitted per treatment arm (section S4.1.3).

S4.1.1 Equivalent model specifications

A logistic heterogeneous treatment effect model as introduced in section 4.4.2 includes both main covariate effects and treatment-covariate interactions. When all covariates (or expansions thereof) in such a model interact with treatment, an exactly equivalent set of 2 models can be specified within the control group and the treated group separately. For instance, a heterogeneous treatment effect model of the form

$$\text{logit}(P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) = \beta_0 + \beta_t a_i + \boldsymbol{\beta_m}^\top \boldsymbol{x}_i + \boldsymbol{\beta_z}^\top \boldsymbol{x}_i a_i \tag{S4.1.1}$$

has a corresponding set of within-treatment group models given by

$$\begin{aligned} \text{logit}(P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x}_i)) &= \beta_0 + \boldsymbol{\beta_m}^\top \boldsymbol{x}_i \\ \text{logit}(P(Y_i = 1|A = 1, \boldsymbol{X} = \boldsymbol{x}_i)) &= (\beta_0 + \beta_t) + (\boldsymbol{\beta_m} + \boldsymbol{\beta_z})^\top \boldsymbol{x}_i \end{aligned} \tag{S4.1.2}$$

Note that these two models are separate models for the potential outcomes of interest (*i.e.* $P(Y^{a=0} = 1|\boldsymbol{X} = \boldsymbol{x_i})$ and $P(Y^{a=1} = 1|\boldsymbol{X} = \boldsymbol{x_i})$ respectively). The other way around, starting from two separate models for both potential outcomes as fitted within each treatment group separately, the models

$$\begin{aligned} \text{logit}(P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x}_i)) &= \beta_{00} + \boldsymbol{\beta}_{\boldsymbol{m0}}^\top \boldsymbol{x}_i \\ \text{logit}(P(Y_i = 1|A = 1, \boldsymbol{X} = \boldsymbol{x}_i)) &= \beta_{01} + \boldsymbol{\beta}_{\boldsymbol{m1}}^\top \boldsymbol{z}_i \end{aligned} \tag{S4.1.3}$$

are equivalent to

$$\begin{aligned} \text{logit}(P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) =& \beta_{00} + (\beta_{01} - \beta_{00})a_i + \\ & \boldsymbol{\beta}_{\boldsymbol{m0}}^\top \boldsymbol{x}_i + (\boldsymbol{\beta_{m1}} - \boldsymbol{\beta_{m0}})^\top \boldsymbol{x}_i a_i \end{aligned} \tag{S4.1.4}$$

The equivalence between these model specifications holds for the maximum likelihood estimates of the $\boldsymbol{\beta}$ parameter vector, but no longer holds when introducing a penalty into the estimation process.

S4.1.2 Penalized maximum likelihood

In case of penalized maximum likelihood, estimates for the separate within treatment-group models will no longer be equivalent to those from a full sample interaction model. For instance, let us consider the case of a ridge or lasso penalty (*i.e.* $\lambda \frac{1}{2} \|\boldsymbol{\beta}\|_2^2$ or $\lambda \|\boldsymbol{\beta}\|_1$ respectively [65]). First, each of the models will have its own estimate of λ, allowing for differences between the within-treatment group models. Second, intercepts are not penalized, and in case of separate models (*e.g.* equation (S4.1.2) and (S4.1.3)), the main treatment effect is retrieved as the different between the two model intercepts (equation (S4.1.4)). Hence, the main treatment effect is penalized by default in the full sample interaction model and is not penalized when using two separate models. Third, in case of ridge regression, the degree of penalization depends on the size of the model coefficients, with larger coefficients being penalized more heavily (due to the square in the penalty term). This is of importance for the treatment-covariate interaction models as specified in equation (S4.1.1) and (S4.1.4), since the expression of the covariate effects under treatment and control conditions is not symmetric in that case (*i.e.* with $\boldsymbol{\beta_m}$ in equation (S4.1.1) reflecting covariate effects under the control condition and $\boldsymbol{\beta_z}$ reflecting changes from $\boldsymbol{\beta_m}$ under the treated condition).

S4.1.3 Simulation results

The simulation settings were exactly the same as in the main text. Full sample HTE models including all treatment-covariate interactions were compared to within-treatment group models including only main effects of the covariates. Models were estimated by means of maximum likelihood, ridge regression, and lasso regression. Figure S4.1 shows the simulation study results with respect to root mean squared prediction error (rMPSE) of the individualized treatment effects. In case of maximum likelihood estimation, the results are of course exactly the same for the different model specifications and are only shown in twofold as a reminder. Lasso treatment-covariate interaction models performed best across all settings. Also, the rMSPE of predicted individualized treatment effects based on ridge treatment-covariate interaction models was generally better than the prediction error for ridge models fitted separately per arm. One exception was ridge regression in large sample size ($N = 3600$), where the per arm models resulted in a better rMSPE. In our simulation settings, which all had variability in coefficient size in the data generating mechanism, the ridge penalty induced clear overshrinkage on large coefficients for all models. This is to be expected due to the square in the penalty and happened in both within-treatment group models and treatment-interaction models. However, in case of treatment-interaction modeling, underfitting of large main effects led to overfitting of the corresponding treatment-covariate interactions [3]. While this happened in all settings and thus across all sample sizes, we hypothesize that the negative effect of this bias on the predictions $\boldsymbol{\delta}$ was offset by more accurate estimation of λ in the full sample treatment-interaction models, except in large sample size settings. Therefore, different model specifications that affect to expected size of the estimated coefficients require careful thought in presence of a ridge penalty. These issues do no affect the lasso penalty. In case of lasso regression, the benefit of having a larger sample size to estimate the penalty parameter λ (*i.e.* as in the treatment-interaction model) led to better performance in all simulation

[3] Note that the square in the ridge penalty means that large estimated coefficients have a larger contribution to the penalty, and are thus more heavily penalized towards zero. An inadvertent characteristic of the treatment-interaction model in case of ridge regression is that the cost of increasing a large main effect parameter (*i.e.* in this context an increase in the effect of the covariate under the control condition), is larger than the cost of the same increase in the smaller corresponding treatment-covariate interaction (*i.e.* the same increase in the effect of the covariate but now under the treatment condition). As a numerical example, assume a main effect coefficient is actually 1 and the corresponding treatment-covariate interaction coefficient is actually 0.5. Shrinking 1 to 0.9 reduces $\|\boldsymbol{\beta}\|_2^2$ by 0.19, and overfitting 0.5 by the same amount increases $\|\boldsymbol{\beta}\|_2^2$ by only 0.11

settings.

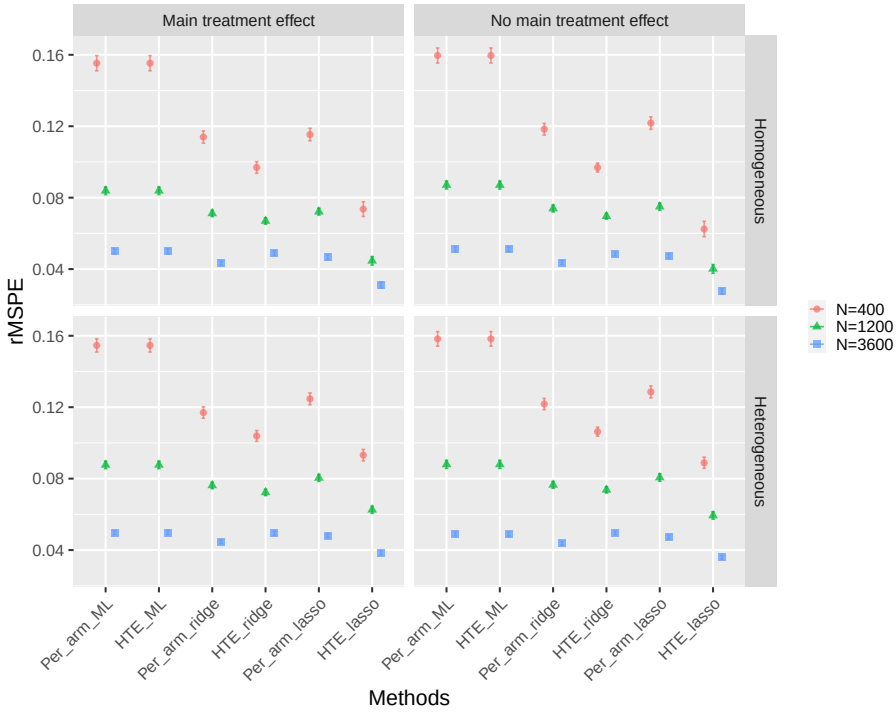


Figure S4.1: Simulation study results: average root mean squared prediction error of the predicted treatment effects (over 250 simulations) with ±2 SE error bars for all simulation settings. Note that the standard errors are often so small that they are obscured by the mean estimates. Abbreviations for the methods are: HTE (heterogeneous treatment effect model), ML (maximum likelihood)

S4.2 Simulating data for a given outcome prevalence

The goal was to simulate data with a prespecified outcome prevalence for the control group. The model underlying the simulations was given in equation (4.7) and is restated here for ease of reference:

$$\text{logit}(P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) = \beta_0 + \beta_t a_i + \boldsymbol{\beta_m}^\top \boldsymbol{x}_i + \boldsymbol{\beta_z}^\top \boldsymbol{z}_i a_i \tag{S4.2.1}$$

For any given treatment condition, this reduces to

$$\text{logit}(P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)) = \beta_{0*} + \boldsymbol{\beta_*}^\top \boldsymbol{x}_i \tag{S4.2.2}$$

where β_{0*} combines β_0 and β_t and $\boldsymbol{\beta_*}$ combines $\boldsymbol{\beta_m}$ and $\boldsymbol{\beta_z}$. Therefore, conditional on treatment condition, the log odds of an event is a linear combination of just the p covariates. Since these had a standard normal distribution by design, their linear combination is also normal with mean equal to β_{0*} and variance equal to

$$\text{Var}(\beta_{0*} + \boldsymbol{\beta_*}^\top \boldsymbol{x}_i) = \boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}^\top = \sigma^2 \tag{S4.2.3}$$

where $\beta = \{\beta_{0*}, \boldsymbol{\beta_*}\}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of the covariates.

Then using

$$\Pr(Y = 1) = \frac{1}{1 + e^{-\beta_{0*} - \sigma Z}} \tag{S4.2.4}$$

where Z is a standard normally distribute random variable, the outcome prevalence or expected probability of $\Pr(Y = 1)$ equals

$$\begin{aligned} \mathbb{E}(\Pr(Y = 1|\boldsymbol{X})) &= \int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{1}{1 + e^{-\beta_{0*} - \sigma Z}} \right) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left(\frac{e^{-z^2/2}}{1 + e^{-\beta_{0*} - \sigma z}} \right) dz \end{aligned} \tag{S4.2.5}$$

Since $\sigma = \sqrt{\boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}^\top}$ only depends on known simulation parameters, the equation can be solved numerically for β_{0*} to get the desired outcome prevalence in a given treatment group.

S4.3 Simulation study calibration results

The `CalibrationFigures.pdf` file (available online as a supplement to DOI: 10.1002/sim.9154) contains calibration plots for $\hat{\delta}(\boldsymbol{x}_i)$, as predicted by each method, versus the true $\delta(\boldsymbol{x}_i)$. Simulation settings with a main treatment effect are denoted as $\beta_t < 0$, settings with a homogeneous treatment effect are denoted as HOM, and settings with a heterogeneous treatment effect as HTE. Each individual plot shows the ideal diagonal in red (with an exception of the absolute null settings where the ideal is $\hat{\delta}(\boldsymbol{x}_i) \equiv 0$). Each black calibration line is the result of a single simulation run and connects the mean predicted $\hat{\delta}(\boldsymbol{x}_i)$ and mean $\delta(\boldsymbol{x}_i)$ in 20 equal-size quantile groups of $\hat{\delta}(\boldsymbol{x}_i)$. The histograms on the x-axis gives an indication of the density of quantile groups over all simulation (the groups vary due to sampling variability).

Chapter 5

Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment

Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment. *arXiv:2209.06101 [stat.ME]*, 2022. (http://arxiv.org/abs/2209.06101) (*under revision*)

Abstract
Personalized medicine constitutes a growing area of research that benefits from the many new developments in statistical learning. A key domain concerns the prediction of individualized treatment effects, and models for this purpose are increasingly common in the published literature. Aiming to facilitate the validation of prediction models for individualized treatment effects, we extend the classical concepts of discrimination and calibration performance to assess causal (rather than associative) prediction models. Working within the potential outcomes framework, we first evaluate properties of existing statistics (including the c-for-benefit) and subsequently propose novel model-based statistics. The main focus is on randomized trials with binary endpoints. We use simulated data to provide insight into the characteristics of discrimination and calibration statistics, and further illustrate all methods in a trial in acute ischemic stroke treatment. Results demonstrate that the proposed model-based statistics had the best characteristics in terms of bias and variance. While resampling methods to adjust for optimism of performance estimates in the development data were effective on average, they had a high variance across replications that limits their accuracy in any particular applied analysis. Thereto, individualized treatment effect models are best validated in external data rather than in the original development sample.

5.1 Introduction

The prediction of individualized treatment effect conditional on patient characteristics has received much interest recently [89, 90, 142, 134, 24, 146]. Such models typically predict a clinically relevant outcome under two different treatment conditions, and the difference between these predictions is attributed to the effect of treatment. This information is of clear interest in the context of clinical decision-making if the underlying model is of sufficient quality. However, the evaluation of individualized treatment effect (ITE) models is still a key methodological challenge and little guidance is currently available on how to quantify their performance [146].

In this paper, we focus on ITE models that contrast the effect of two treatment conditions on the risk of a binary endpoint. More specifically, we focus on assessment of their performance; guidance on their development is available elsewhere (*e.g.*, [90, 146, 113]). Typical measures of prediction model performance with respect to *outcome risk* predictions include measures of calibration and discrimination [9, 8, 147]. However, our specific interest here is in predictions of *risk difference* attributed to the effect of treatment (*i.e.*, in absolute individualized treatment effect predictions). Although calibration and discrimination performance can also be assessed at the *risk difference* (treatment effect) level, existing measures (*e.g.*, calibration intercept, calibration slope, c-statistic) do not apply without modification because individual treatment effects (in contrast to regular outcomes) are unobservable [146]. For this reason, a new c-statistic was recently proposed that applies to absolute treatment effect predictions in settings with a binary endpoint, along with a quantile-based assessment of calibration [132].

We expand on this previous work by casting the entire prediction and evaluation process in the potential outcomes framework [94, 93] and by developing model-based measures of discrimination and calibration performance with respect to individualized treatment effect predictions. Herein, the potential outcomes framework provides a way to deepen understanding of what is actually being measured. The model-based measures make more efficient use of the data without relying on matching on arbitrary cut-offs.

Section 5.2 sets the scene and describes the challenge of individualized causal prediction in terms of the potential outcomes framework. Subsequently, Section 5.3 and Section 5.4 describe existing and novel measures of discrimination and calibration with respect to absolute treatment effect respectively. Simulation results are provided for illustrative purposes. An applied example using data

from the third International Stroke Trial (IST-3) [148] is described in Section 5.6. Lastly, Section 5.8 provides a general discussion.

5.2 Individualized treatment effect prediction

Most outcome prediction research focuses on capturing statistical association in absence of interventions. Individualized treatment effect (ITE) prediction is a different type of prediction since it has a causal interpretation: the quantity to be predicted is the effect caused by the treatment (or intervention, in a larger sense) on the outcome. Therefore, before moving to the performance measures of interest, this section shortly outlines causal prediction. Subsequently, issues surrounding the use of binomial outcome data for ITE modeling are shortly discussed (further details are available as online supplementary material S5.1).

5.2.1 Causal prediction

To emphasize the causal nature of the predictions, it is helpful to write the individualized treatment effect of interest in terms of the potential outcomes framework [94, 93]. For treatment taking values $a \in \mathcal{A}$, $Y^{A=a}$ denotes the potential outcome under treatment a. When comparing two treatments, the ITE for individual $i, \ldots, n$ can be defined as

$$\delta(\boldsymbol{x}_i) = \mathbb{E}(Y_i^{a=1}|\boldsymbol{X} = \boldsymbol{x}_i) - \mathbb{E}(Y_i^{a=0}|\boldsymbol{X}_i = \boldsymbol{x}_i) \tag{5.2.1}$$

where $\boldsymbol{x}_i$ is a row vector of individual-level characteristics in matrix $\boldsymbol{X}$. The degree of granularity or individualization reflected by $\delta(\boldsymbol{x}_i)$ relates to the number of predictors included in $\boldsymbol{X}$, to the strength and shape of their association with the potential outcomes, and especially to the degree to which they have a differential effect across potential outcomes (*i.e.*, modify the effect of treatment). Ideally, the set of measured individual-level characteristics includes all relevant characteristics with respect to individualized treatment effect. In practice however, this set of all relevant characteristics is often unknown and the best way forward is to aim for conditioning on the most important characteristics. Correspondingly, equation (5.2.1) reflects ITE as a conditional treatment effect for some set of characteristics.

Since in practice only one potential outcome is observed per individual [23], assumptions are required to estimate $\delta(\boldsymbol{x}_i)$ based on the observed data. These assumptions are discussed in detail elsewhere [146, 22]. In short, the key assumptions are *exchangeability* (the potential outcomes do not depend on the

assigned treatment), *consistency* (the observed outcome under treatment $a \in \mathcal{A}$ corresponds to the potential outcomes $Y^{A=a}$), and positivity (each individual has a non-zero probability of each treatment assignment. An additional assumption that eases inference is *no interference* (the potential outcomes for individual i do not depend on treatment assignment to other individuals). Based on these assumptions, the individualized treatment effect can be identified given the observed data:

$$\begin{aligned}\delta(\boldsymbol{x}_i) &= \mathbb{E}(Y_i^{a=1}|\boldsymbol{X} = \boldsymbol{x_i}) - \mathbb{E}(Y_i^{a=0}|\boldsymbol{X} = \boldsymbol{x_i}) \\ &= \mathbb{E}(Y_i^{a=1}|A = 1, \boldsymbol{X} = \boldsymbol{x_i}) - \mathbb{E}(Y_i^{a=0}|A = 0, \boldsymbol{X} = \boldsymbol{x_i}) \\ &\quad \text{(by exchangeability)} \\ &= \mathbb{E}(Y_i|A = 1, \boldsymbol{X} = \boldsymbol{x_i}) - \mathbb{E}(Y_i|A = 0, \boldsymbol{X} = \boldsymbol{x_i}) \\ &\quad \text{(by consistency)} \end{aligned} \tag{5.2.2}$$

Equation (5.2.2) shows that ITE predictions ($\hat{\delta}(\boldsymbol{x}_i)$) can be estimated using a prediction model for outcome risk $\mathbb{E}(Y_i|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)$. Many modeling tools can be used for this endeavor and the details are beyond the scope of this paper and are given elsewhere (*e.g.,* [146, 135]).

5.2.2 Binary outcome data

Focusing on binary outcomes, we observe outcome $Y_i \in \{0, 1\}$ and covariate status $\boldsymbol{x}_i$ for each individual i. In this context, the ITE estimate $\delta(\boldsymbol{x}_i)$ is a difference between two risk predictions ($P(Y_i = 1|A = 1, \boldsymbol{X} = \boldsymbol{x}_i) - P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x}_i)$). The range of $\hat{\delta}(\boldsymbol{x}_i)$ includes all values in the $[-1, 1]$ interval, while the observed difference between any two outcomes can only be one of $\{-1, 0, 1\}$. Therefore, in addition to the challenge that each individual only has an observed outcome for one treatment condition, the observations come with large and irreducible binomial error and hence provide only limited information. A further consideration with predictions for binary outcome data is that they are commonly non-linear functions of the covariates, and hence the effects of treatment and the covariates are usually not additive on the risk difference scale of interest here. Consequently, the resulting ITE predictions conflate variability from different sources: between-subject variability in $P(Y^{a=0}|X = x)$ and genuine treatment effect heterogeneity on the scale used for modeling. This is the price to pay for the benefit in terms of interpretation of measures on the scale of $\delta(x)$ [92].

5.2.3 The challenge

In practice, the fundamentally limited nature of observed data when it comes to causal inference (*i.e.*, with only one potential outcome being observed), the irreducible binomial error affecting the risk difference twice, and the challenge of model specification and estimation are all present at the same time. This evidently poses a challenge at the time of model development, but within the scope of this paper, it certainly also poses challenges during evaluation of models predicting individualized treatment effects. Most notably, and in contrast with regular prediction modeling, a direct comparison between predictions and observed outcomes is not feasible.

In this paper, we evaluate discrimination and calibration at the level of predicted individualized treatment effects. To this end, we first evaluate a recently introduced metric to assess discriminative performance [132] and subsequently propose alternative procedures that aim to alleviate some of the shortcomings. Thereafter, we address calibration of predicted treatment effect. With respect to the detection of overfitting (*i.e.*, overly complex models that fail to generalize), we examine performance in both internal and external validation settings.

5.3 Discrimination for individualized treatment effects

Discriminative model performance reflects the degree to which model predictions are correctly rank-ordered and is a common performance measure in regular prediction modeling [8, 9]. In the context of outcome risk prediction, the observed outcomes provide an immediate reference to check rank-ordering at the level of the predictions. However, such a direct reference is not available for ITE models since individual treatment effects cannot be observed directly, which necessitates approximations. One of the possibilities is to use matching and is used in a recent proposal, the c-for-benefit, for a measure of discriminative performance on the ITE level [132]. The section shortly outlines the c-for-benefit and subsequently discusses its properties, limitations, and possible extensions.

5.3.1 C-for-benefit definition

In the setting of a two-arm study measuring a binary outcome of interest, the c-for-benefit aims to assess discrimination at the level of ITE predictions (re-

ferred to as 'predicted [treatment] benefit' in the original paper).[1] The problem of unobserved individual treatment effects is approached from a matching perspective. One-to-one matching is used to match treated individuals to control individuals based on their predicted treatment effects. The subsequent data pairs hence consist of a treated individual and a control individual with similar predicted treatment effect. Observed treatment effect in the pair is defined as the difference in outcomes between these two individuals. Of note, observed (within-pair) treatment effect can only be {-1,0,1}. Subsequently, c-for-benefit has been defined as "the proportion of all possible pairs of matched individual pairs with unequal observed benefit in which the individual pair receiving greater treatment benefit was predicted to do so" [132]. The predicted treatment effect within each pair used in this definition is taken to be the (within-pair) average of predicted treatment effects. That is, for a pair comprising control individual i out of $1, \ldots, n_i$ and treated individual j out of $1, \ldots, n_j$, predicted treatment effects are taken to be

$$\hat{\delta}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \{(\hat{P}(Y_i|A_i = 1, \boldsymbol{X} = \boldsymbol{x}_i) - \hat{P}(Y_i|A_i = 0, \boldsymbol{X} = \boldsymbol{x}_i)) + (\hat{P}(Y_j|A_j = 1, \boldsymbol{X} = \boldsymbol{x}_j) - \hat{P}(Y_j|A_j = 0, \boldsymbol{X} = \boldsymbol{x}_j))\}/2 \quad (5.3.1)$$

The 'observed' treatment effect is subsequently taken to be $O_{ij} = Y_i - Y_j$. Therefore, the c-for-benefit is a regular concordance statistic (c-statistic) as commonly applied in survival data [149, 8], but applied to pairs of individuals that underwent different treatments. If the two (binary) outcomes in such a pair are discordant, then there supposedly is some evidence of a treatment effect (*i.e.*, benefit or harm); conversely, there is no such evidence when the outcomes are concordant (*i.e.*, the predicted treatment effect did not manifest as a difference in outcomes). The implicit assumption is that individual i and j are similar enough to serve as pseudo-observations of the unobserved potential outcomes. In the ideal case where $\boldsymbol{x}_i = \boldsymbol{x}_j$ this is indeed the case, but such perfect matches are unlikely to be available for multivariable prediction models.[2] An alternative (unsupervised) matching procedure that was proposed in the same paper is to

[1] The original paper did not focus on the required conditions for *causal* interpretation of the predicted individualized treatment effects; here we assume that these assumptions, as described in Section 5.2.1, are met.

[2] Note that we here forgo the notion of including *all* relevant covariates, since $\boldsymbol{x}_i = \boldsymbol{x}_j$ is sufficient for the degree of individualization reflected by the model.

match on covariates in terms of Mahalanobis distance.[3] For the remainder of this paper, we will use cben-$\hat{\delta}$ to refer to the original c-for-benefit using 1:1 matching on predicted treatment effect.

5.3.2 C-for-benefit challenges

Although the c-for-benefit has been applied on several occasions (*e.g.*, [150, 151, 152]), its properties have not been fully elucidated. Van Klaveren et al. [132] recommended further work on its theoretical basis and simulations studies, which we here present. Evidently, many issues that apply to the regular concordance statistic also apply to the c-for-benefit. However, since the c-for-benefit relates to risk differences and depends on outcomes that cannot be observed directly, additional challenges arise which we outline below.

Difficult interpretation

As described, the c-for-benefit uses 1:1 matching *and* averages ITEs within each pair of matched individuals (*i.e.*, $\hat{\delta}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in equation (5.3.1)). As we will see below (section 5.3.3), this average of two ITEs does not generally correspond to the treatment effect induced by the study design even if the model is correctly specified. Also, the observed outcome difference O_{ij} reflects more than just $\hat{\delta}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ unless both control outcome risk and treated outcome risk are the same for matched individuals. These two issues obfuscate the interpretation of the index.

Sensitivity to matching procedure

Two matching procedures were proposed for the c-for-benefit: i) based on $\hat{\boldsymbol{\delta}}$ (*i.e.*, minimize the distance between pairs $\hat{\delta}_i$ and $\hat{\delta}_j$), and ii) based on the Mahalanobis distance between covariate vectors[3] [132]. In theory, matching on covariates $\boldsymbol{X}$ leads to appropriate matches on predicted treatment effects since the latter is a function of the covariates. However, the reverse is not true: matching on $\hat{\boldsymbol{\delta}}$ does not necessarily lead to appropriate matches on $\boldsymbol{X}$. The reason is that multiple configurations of $\boldsymbol{X}$ can give rise to the same value of $\hat{\delta}$, which does not satisfy equation (5.2.2). Importantly, this is even the case for a correctly specified model. The only setting in which matching on predicted ITEs is guaranteed to generate appropriate matches on $\boldsymbol{X}$ is when $\hat{\boldsymbol{\delta}}$ is a bijective function of $\boldsymbol{X}$ (*i.e.*,

[3] where the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined as $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)' \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \boldsymbol{x}_j)}$ with $\boldsymbol{S}$ the covariance matrix of the covariates in $\boldsymbol{X}$

$\hat{\boldsymbol{\delta}}$ and $\boldsymbol{X}$ have a one-to-one correspondence). However, this is highly atypical in prediction modeling (*e.g.*, a model with only one covariate that has a functional form with a strictly positive or negative first derivative). Also, both matching procedures were proposed for $1:1$ matching, which requires either equal groups size for both study arms or loss of data. A simple remedy that stays close to the original idea is to perform repeated analysis with random sub-samples of the larger arm. Alternatively, the implementation of many-to-one matching (*e.g.*, full matching) or many-to-many matching [153, 154, 155] might be implemented, but none of these has been studied in the context of the c-for-benefit.

5.3.3 Towards a more principled concordance statistic for benefit

The c-for-benefit compares concordance between differences in i) average predicted treatment effect within matched control-treated pairs $\hat{\delta}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and ii) observed outcome differences within those same pairs O_{ij}. However, in general $\delta_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) \neq \mathbb{E}(O_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ unless $\boldsymbol{x}_i = \boldsymbol{x}_j$. This section decomposes $\mathbb{E}(O_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$ to find conditions under which unbiased comparison to ITE predictions is available. Thereto, for controls $i \in 1, \ldots, n_i$ and treated individuals $j \in 1, \ldots, n_j$ and writing $g_0(\boldsymbol{x})$ for $P(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x})$ and $g_1(\boldsymbol{x})$ for $P(Y_i = 1|A = 1, \boldsymbol{X} = \boldsymbol{x})$,

$$
\begin{aligned}
\mathbb{E}(O_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j) =& \mathbb{E}(Y_j|\boldsymbol{x}_j - Y_i|\boldsymbol{x}_i) \\
=& \mathbb{E}(Y_j|\boldsymbol{x}_j) - \mathbb{E}(Y_i|\boldsymbol{x}_i) \\
=& g_1(\boldsymbol{x}_j) - g_0(\boldsymbol{x}_i) && (5.3.2) \\
=& [g_0(\boldsymbol{x}_j) + \delta(\boldsymbol{x}_j)] - g_0(\boldsymbol{x}_i) && (5.3.3) \\
=& g_1(\boldsymbol{x}_j) - [g_1(\boldsymbol{x}_i) - \delta(\boldsymbol{x}_i)] && (5.3.4)
\end{aligned}
$$

Hence, from equation (5.3.3) and given $g_0(\cdot)$, the expected observed outcome difference between individual j receiving treatment and individual i receiving control equals the true equals the true individualized treatment effect for individuals sharing the same characteristics $\boldsymbol{x}$ as j

$$
\mathbb{E}(O_{ij}|g_0(\boldsymbol{x}_i), g_0(\boldsymbol{x}_j)) = \mathbb{E}(Y_j - g_0(\boldsymbol{x}_j)) - \underbrace{\mathbb{E}(Y_i - g_0(\boldsymbol{x}_i))}_{0} = \delta(\boldsymbol{x}_j), \qquad (5.3.5)
$$

and analogously, from equation (5.3.4) and given $g_1(\cdot)$, the expected observed outcome difference between individual j receiving treatment and individual i

receiving control equals the true individualized treatment effect for individuals sharing the same characteristics $\boldsymbol{x}$ as i

$$\mathbb{E}(O_{ij}|g_1(\boldsymbol{x}_i), g_1(\boldsymbol{x}_j)) = \underbrace{\mathbb{E}(Y_j - g_1(\boldsymbol{x}_j))}_{0} - \mathbb{E}(Y_i - g_1(\boldsymbol{x}_i)) = \delta(\boldsymbol{x}_i) \qquad (5.3.6)$$

Conditioning on $g_0(\cdot)$ in equation (5.3.5) aims to achieve prognostic balance, which bears resemblance to prognostic score analysis [156, 157]. Conditioning on $g_1(\cdot)$ in equation (5.3.6) is just the mirror image for $g_1(\cdot)$. In practice, $g_0(\cdot)$ and/or $g_1(\cdot)$ will of course have to be estimated and the exact equalities will become approximations. For continuous outcomes, equation (5.3.5) allows evaluation of predictions $\hat{\delta}(\boldsymbol{x}_j)$ against residuals $Y_j - \hat{g}_0(\boldsymbol{x}_j)$ for $j = 1 \ldots, n_j$, and equation (5.3.6) allows evaluation of predictions $-\hat{\delta}(\boldsymbol{x}_i)$ against residuals $Y_i - \hat{g}_1(\boldsymbol{x}_i)$ for $i = 1 \ldots, n_i$.[4] A key benefit of this approach is that matching is not required. However, extension of such a residual-based approach to binary outcome data is not clear. Hence, we implemented a 1:1 matching procedure similar to cben-$\hat{\delta}$, but with two important differences. First, matching was performed based on $\hat{g}_0(\boldsymbol{x})$ as opposed to predicted treatment effect. Thereby, whenever $\hat{g}_0(\boldsymbol{x}_i) = \hat{g}_0(\boldsymbol{x}_j)$, the expected difference between Y_i and Y_j *does* equal $\hat{\delta}(\boldsymbol{x}_j)$ if the models for g_0 and g_1 are correct. Second, and following from the previous, this implementation evaluated concordance between $\hat{\delta}(\boldsymbol{x}_j)$ (as opposed to $\hat{\delta}_{ij}$) and the corresponding O_{ij}'s. We will further refer to this implementation as cben-$\hat{y}^0$. Note that a mirror image alternative could be performed when matching on $\hat{g}_1(\boldsymbol{x})$; the choice between the two might be guided by the expected quality in terms of prediction accuracy of $\hat{g}_0(\cdot)$ and $\hat{g}_1(\cdot)$, and the size of the group in which ITE predictions will be evaluated.

5.3.4 Model-based c-statistics for individualized treatment effect

Extending earlier work on model-based concordance assessment in the context of risk prediction [158], we propose model-based concordance assessment on the level of absolute individualized treatment effect prediction. The concordance probability that we aim for is the probability that any randomly selected pair of patients (regardless of treatment assignment) has concordant ITE predictions

[4] Either way, one arm can be used to estimate the relevant $\hat{g}(\cdot)$ and the other arm to evaluate $\hat{\delta}(\cdot)$. Alternatively, an external model $\hat{g}(\cdot)$ might be used, but, as demonstrated in the simulation study, bias will be introduced if this external model does not fit the data well.

and outcomes, divided by the probability that their outcomes are different. While the ITE predictions are clearly defined (equation (5.2.2)), the outcomes (individualized treatment effects) are never observed directly and can be approximated in multiple ways. The model-based approach is to use the model's predictions of the potential outcomes when deriving the concordance statistic. Therefore, it reflects the concordance statistic that would be expected under the assumption that the model is correct and given a specific set of data. Note that all information required for a model-based estimate is in the model, so there is no problem with respect to unobserved potential outcomes.

Let us first define concordance between ITE predictions and potential outcome patterns in line with the c-for-benefit. As above, take an event Y=1 to be harmful. For a randomly selected individual k with a lower predicted ITE than another individual l ($\hat{\delta}_k < \hat{\delta}_l$, where $k, l \in 1, \ldots, n$ and $k \neq l$), treatment is predicted to be more beneficial (or less harmful) for individual k as compared to individual l. The potential outcome patterns that are concordant with $\hat{\delta}_k < \hat{\delta}_l$ are

1. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 0, Y_l^{a=0} = 0$ (benefit for k, no benefit for l)
2. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 1$ (benefit for k, no benefit for l)
3. $Y_k^{a=1} = 0, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (benefit for k, harm for l)
4. $Y_k^{a=1} = 0, Y_k^{a=0} = 0, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (no benefit for k, harm for l)
5. $Y_k^{a=1} = 1, Y_k^{a=0} = 1, Y_l^{a=1} = 1, Y_l^{a=0} = 0$ (no benefit for k, harm for l).

The corresponding estimated probabilities of these patterns follow easily from the model(s) for both potential outcomes. For instance, for the first pattern: $[1 - \hat{P}(Y_k^{a=1} = 1)] \cdot \hat{P}(Y_k^{a=0} = 1) \cdot [1 - \hat{P}(Y_l^{a=1} = 1)] \cdot [1 - \hat{P}(Y_l^{a=0} = 1)]$. The sum of the five patterns is further referred to as $P_{\text{benefit},k,l}$. Likewise, let $P_{\text{harm},k,l}$ denote the total probability of observing relative harm for case k with respect to case l, which can be obtained in a similar manner. Returning to our definition of concordance probability, the estimated probability of concordant ITE predictions and potential outcomes for two randomly chosen patients k and l for any given model can be written as

$$\hat{P}(\text{concordant}) = \frac{1}{n(n-1)} \sum_k \sum_{l \neq k} \left[I(\hat{\delta}_k < \hat{\delta}_l) \hat{P}_{\text{benefit},k,l} + I(\hat{\delta}_k > \hat{\delta}_l) \hat{P}_{\text{harm},k,l} \right] \tag{5.3.7}$$

Subsequently, the model-based probability estimate of a potential outcome pattern reflecting either relative benefit or relative harm for a randomly selected pair of patients is

$$\frac{1}{n(n-1)} \sum_k \sum_{l \neq k} \left[\hat{P}_{\text{benefit},k,l} + \hat{P}_{\text{harm},k,l} \right], \tag{5.3.8}$$

and hence the concordance probability is

$$\frac{\sum_k \sum_{l \neq k} \left[I(\hat{\delta}_k < \hat{\delta}_l) \hat{P}_{\text{benefit},k,l} + I(\hat{\delta}_k > \hat{\delta}_l) \hat{P}_{\text{harm},k,l} \right]}{\sum_k \sum_{l \neq k} \left[\hat{P}_{\text{benefit},k,l} + \hat{P}_{\text{harm},k,l} \right]} \tag{5.3.9}$$

A formulation that allows for ties $\hat{\delta}_k = \hat{\delta}_l$ and avoids the need to derive $\hat{P}_{\text{harm},k,l}$ is, in line with the Harrell's c-statistic [149, 8],

$$\text{mbcb} = \frac{\sum_k \sum_{l \neq k} \left[I(\hat{\delta}_k < \hat{\delta}_l) \hat{P}_{\text{benefit},k,l} + \frac{1}{2} I(\hat{\delta}_k = \hat{\delta}_l) \hat{P}_{\text{benefit},k,l} \right]}{\sum_k \sum_{l \neq k} \left[\hat{P}_{\text{benefit},k,l} \right]} \tag{5.3.10}$$

We propose the model-based c-for-benefit (mbcb) as a model-based alternative to the c-for-benefit, hence the name. Estimating both $\hat{\delta}(\boldsymbol{x})$ and $\hat{P}_{\text{benefit},k,l}$ from the same model, the mbcb provides the theoretical concordance probability between ITE predictions and potential outcomes that would be achieved if the model is correct. Important to note, such a model-based statistic only depends on the observed outcomes in the development data through the model, and hence does not provide any insight into model fit. For instance, estimating both $\hat{\delta}(\boldsymbol{x})$ and $\hat{P}_{\text{benefit},k,l}$ from some ITE model $\hat{\delta}_m$ in new data D, the mbcb would return the expected concordance probability for $\hat{\delta}_m$ at the ITE-level as based on the distribution of $\boldsymbol{X}$ in D, and assuming $\hat{\delta}_m$ is correct; it does not depend on the outcomes measured in D. In other words, it provides a case-mix adjusted (*i.e.*, adjusted for the sampled $\boldsymbol{X}$) expected mbcb for new data [158]. This is of interest since concordance statistics are known to be sensitive to case-mix. For instance, discriminative performance in terms of a concordance statistic for a new sample that is truncated in terms of $\boldsymbol{X}$ (*e.g.*, due to inclusion criteria) will be lower, even if the model is perfectly adequate, just because it is harder to discriminate in the new sample [159]. Hence, the case-mix adjusted expected mbcb is a better reference than the mbcb in the development data when validating a model. To obtain a validation estimate of ITE-level concordance

probability (*i.e.*, without assuming that the ITE model is correct), the estimator for $\hat{P}_{\text{benefit},k,l}$ should be based on independent validation data that were not used for ITE model development. The main goal is to obtain estimates of $\hat{P}_{\text{benefit},k,l}$ as accurate as possible for the new data, since these take the role of the 'observed' outcomes for the model-based c-for-benefit. A way to do so it by refitting the original ITE model in the independent data and using the resulting outcome risk predictions to calculate $\hat{P}_{\text{benefit},k,l}$.

5.4 Calibration of individualized treatment effect predictions

A calibration measure reflects the degree to which predictions (predicted treatment effect) agree with observations (observed treatment effect). Several routes can be taken when interest is in predicted individualized treatment effect.

(A) Classical calibration: compare predicted outcome risk under the assigned treatment conditions versus observed outcomes.

(B) Within-arm classical calibration: classical calibration within treatment arms.

(C) Quantile-group calibration of average individualized versus observed treatment effect.

(D) Model-based calibration of individualized versus observed treatment effect.

Method (A) has been described at length in the literature (*e.g.*, [9, 8, 147]) and method (B) is a straightforward extension. Both have the disadvantage that they do not directly assess absolute treatment effect: overall calibration of outcome risk may look good when prognostic factors are well modeled and explain most of the outcome risk, even though a comparatively small (but possibly important) treatment effect is not well represented.

A common way to proceed in the direction of direct predicted treatment effect evaluation is to form quantile groups of the predictions and to compare (average) predicted and observed treatment effect within these groups [132, 146] (method ((C))). However, the cut-off points to form these groups are always arbitrary and smooth model-based calibration plots have become the preferred method of choice in regular (outcome risk) calibration assessment [147]. This leaves

method (D) which, in theory, provides the desired direct ITE assessment while avoiding the disadvantages associated with cut-offs. However, to our knowledge, such a method has not been described yet. We propose a model-based approach that isolates the calibration of $\hat{\delta}$ based on equation (5.3.5).

According to equation (5.3.5), $\mathbb{E}(Y_j - \hat{g}_0(\boldsymbol{x}_j))$ can be directly compared against predictions $\hat{\delta}(\boldsymbol{x}_j)$ for treated individuals $j \in 1, \ldots, n_j$. In case of dichotomous Y_j, a natural way to do so is to model Y_j with offset $\hat{g}_0(\boldsymbol{x}_j)$. For instance, for a logistic model, a calibration model could be formulated as

$$\text{logit}(Y_j) = \beta_0 + \beta_1 \hat{\delta}_{lp}(\boldsymbol{x}_j) + \hat{g}_{lp,0}(\boldsymbol{x}_j) \tag{5.4.1}$$

where $\hat{\delta}_{lp}(\boldsymbol{x}_j) = \text{logit}(\hat{g}_1(\boldsymbol{x}_j)) - \text{logit}(\hat{g}_0(\boldsymbol{x}_j))$ and $\hat{g}_{lp,0}(\boldsymbol{x}_j) = \text{logit}(\hat{g}_0(\boldsymbol{x}_j))$. The anticipated intercept β_0 and slope β_1 in case of a perfect prediction are 0 and 1 respectively, as for regular prognostic model calibration [9, 8]. Assuming that $\hat{g}_0(\boldsymbol{x}_j)$ is correct, the estimated slope $\hat{\beta}_1$ directly reflects ITE overfitting (slope below 1) or underfitting (slope above 1), and the estimated intercept $\hat{\beta}_0$ reflects average error in the ITE predictions. When $\hat{g}_0(\boldsymbol{x}_j)$ is misspecified, $\hat{\beta}_0$ and $\hat{\beta}_0$ amalgamate ITE calibration and errors in $\hat{g}_{lp,0}(\boldsymbol{x}_j)$. Given the importance of $\hat{g}_0(\boldsymbol{x}_j)$, which essentially anchors the ITE predictions, it might be preferable to derive predictions $\hat{g}_0$ based on a new model fitted in the external control arm data to reduce bias in the assessment of ITE calibration. A more direct way to assess average error in predicted ITEs is to examine the difference between observed and expected average treatment effect: $[\frac{1}{n_j}\sum_j Y_j - \frac{1}{n_i}\sum_i Y_i] - [\frac{1}{n_j}\sum_j \hat{g}_1(\boldsymbol{x}_j) - \frac{1}{n_i}\sum_i \hat{g}_0(\boldsymbol{x}_i)]$.

As a side note, continuous outcomes Y_j allow for direct analysis of residuals $Y_j - \hat{g}_0(\boldsymbol{x}_j)$ by means of a linear regression model.

$$Y_j - \hat{g}_0(\boldsymbol{x}_j) = \beta_0 + \beta_1 \hat{\delta}(\boldsymbol{x}_j) + \epsilon_j \tag{5.4.2}$$

for individuals $j \in 1, \ldots, n_j$ and with $\epsilon_j \sim N(0, \sigma^2)$. The anticipated intercept, slope, and procedure to derive calibration in the large are the same as for (5.4.1). In addition to model-based evaluation, a smooth curve such as a loess (locally estimated scatterplot smoothing) estimate can be drawn through a scatterplot of $Y_j - \hat{g}_0(\boldsymbol{x}_j))$ versus $\delta(\boldsymbol{x}_j)$ to provide a visual evaluation of ITE calibration for continuous outcomes.

5.5 Simulation study

A simulation study was performed with the aim to compare performance of the different discrimination and calibration measures for ITE predictions discussed across varying sample sizes. The simulation study was performed and reported in line with recommendations by Morris et al. [124] and using `R` statistical software version 4.2 [160].

5.5.1 Simulation study procedures

Data generating mechanisms: Synthetic trial data were simulated for a trial comparing two treatments on a binary outcome. Covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ were generated from independent standard normal distributions and treatment assignment was 1:1 and independent of $\boldsymbol{X}$. Data were simulated for both potential outcomes based on a logistic data generating mechanism (DGM) according to model

$$\text{logit}(P(Y_i^{A=a} = 1)) = -1 - 0.75a_i + x_{i1} + 0.5a_i x_{i2} \tag{5.5.1}$$

for a population of size 100,000, and will be further referred to as DGM-1. DGM-1 includes main effects of treatment and X_1 and an interaction between treatment and X_2. For each of $n_{sim} = 500$ simulation runs, development (D) and validation data (V1) sets of size 500, 750, and 1000 were randomly drawn from the population. Marginal event probabilities were $P(Y^{a=0}) \approx 0.31$ and $P(Y^{a=1}) \approx 0.20$. Additionally, independent validation sets of 1000 cases (V2) were sampled from a population of size 100,000 generated from a second DGM (DGM-2) with changes in the coefficients to reflect a different population

$$\text{logit}(P(Y_i^{A=a} = 1)) = -0.5 - 0.5a_i + 0.75x_{i1} + 0.25x_{i2} + 0.25a_i x_{i1} + 0.25a_i x_{i2}, \tag{5.5.2}$$

Marginal event probabilities for the second DGM were $P(Y^{a=0}) \approx 0.39$ and $P(Y^{a=1}) \approx 0.31$. With differences in both average treatment effect and heterogeneity of treatment effect between DGM-1 and DGM-2, a model developed in a sample from DGM-1 should not perform well in individuals from DGM-2.

Estimands: For discrimination, our estimand θ_d is the concordance statistic between ITE predictions and the true probabilities to observe benefit. For a fixed ITE model, a given data generating mechanism, and a fixed matrix of observed covariates, this is exactly the definition of the mbcb in equation (5.3.10)

when substituting true the value of $P_{\text{benefit},k,l}$ (know from the DGM) for the estimated $\hat{P}_{\text{benefit},k,l}$. This provides the expected ITE concordance statistic with the expectation taken over repeated samples of the potential outcomes. Due to its dependence on the matrix of observed covariates in a sample, it is further referred to as the 'sample reference'. For the estimand in the population, the mbcb is a considerable computational burden due to the fast-growing number of observation pairs with increasing sample size. Instead, the 'population reference' was not based on the expectation over potential outcomes given the covariates, but on a single sample of the potential outcomes. Hence it is still unbiased with respect to the true estimand in the population. Further details are provided in online supplementary material S5.2) and also show the relation between the mbcb and Harrell's c-statistics [149, 8] applied to ITE predictions and simulations of both potential outcomes for each individual.

For calibration performance, our estimands were the calibration intercept β_0 and calibration slope β_1 as defined in equation (5.4.1). The true values follow directly from equation (5.4.1) when taking, for $j \in 1, \ldots, n_j$, the known probabilities $P(Y_j^{a=1})$ based on the appropriate DGM, $\hat{\delta}_{lp}(\boldsymbol{x}_j)$ as the sample-based ITE predictions under evaluation, and $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ based on known probabilities $P(Y_j^{a=0})$. Both a sample reference (the value of the estimand for the distribution of $\boldsymbol{X}$ in the given sample) and a population reference (the value of the estimand for the distribution of $\boldsymbol{X}$ in the given population) were derived.

Methods: The ITE model fitted to the development data was a logistic regression model estimated by means of maximum likelihood of the form

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 a_i + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 a_i x_{i1} + \beta_5 a_i x_{i2} \qquad (5.5.3)$$

Discrimination performance was assessed by means of the c-for-benefit using 1:1 benefit matching (cben-$\hat{\delta}$), the c-for-benefit using 1:1 matching on predicted outcome risk under the control treatment (cben-$\hat{y}^0$), and the mbcb. Calibration performance was assessed according to equation (5.4.1). Each of the performance measures was evaluated (1) without correction in the same samples as in which the ITE model was developed (sample D, apparent performance[8]), (2) in interval validation using bootstrap 0.632+ adjustment (3) in interval validation using bootstrap optimism correction, (4) in external validation samples V1, (5) in external validation data samples V2, (6) in the external population from DGM-1, and (7) in the external population from DGM-2. A more detailed account of the procedures in available in online supplementary material S5.3.

Performance measures: Writing θ_s for the reference value in simulation run

s, and $\hat{\theta}_s$ for the corresponding estimate, performance measures were averaged across simulations $s \in 1, \ldots, n_{sim}$ in terms of bias $\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\theta_s - \hat{\theta}_s)$ and root mean squared prediction error $\sqrt{\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\theta_s - \hat{\theta}_s)^2}$.

Additional reference: For calibration evaluation only, a 'naive' population reference value was derived for each ITE model to demonstrate that ITE calibration heavily relies on the accuracy of control outcome risk predictions ($\hat{g}_0$). This reference does not correspond to an estimand of interest, but instead corresponds to 'naive' adjustment for $\hat{g}_0$ as predicted by the evaluated ITE model (*i.e.*, instead of $\hat{g}_0$ as predicted from a local model in data independent from the development data). Thereby, it serves to illustrate the large-sample error that occurs under misspecification of the model for $\hat{g}_0$.

5.5.2 Discrimination results

Figure 5.1 and Table 5.1 show the main simulation results with respect to the discrimination statistics. Tabulated results corresponding to Figure 5.1 are available as online supplementary Table S5.1. First, note that the sample reference and population reference show near perfect agreement across all panels. This shows that the estimand in a validation sample generalized well to entire population (*i.e.*, did not greatly depend on the specific sample of covariate values in a given validation sample).

With respect to the estimates, and starting with apparent evaluations (top left), all statistics showed optimism with respect to the reference standards, which decreased with increasing sample size. As expected, direct evaluation in new data from the same DGM (top-middle) removes optimism for cben-$\hat{\delta}$ and cben-$\hat{y}^0$, and *did not* remove optimism in the model-based c-for-benefit. The latter preserves overfitting since it only estimates the c-statistic that would be obtained for the new data if the model were correct. Note that the estimated cben-$\hat{\delta}$ in V1 was actually too low, indicating bias in the estimator.

As shown in the bootstrap panels in Figure 5.1, both types of bootstrap evaluations adjusted for optimism in apparent evaluations. On average, bias was almost eliminated from cben-$\hat{y}^0$ and the mbcb. For cben-$\hat{\delta}$ the bootstrap adjusted estimates were too low, which is in line the findings in V1. Nonetheless, bootstrap procedures were generally not able to decease the root mean squared prediction error between the estimated statistic and population reference statistic (Table 5.1). The 0.632+ procedure for the cben-$\hat{y}^0$ forms an exception, decreasing both bias and rmse for all sample sizes.

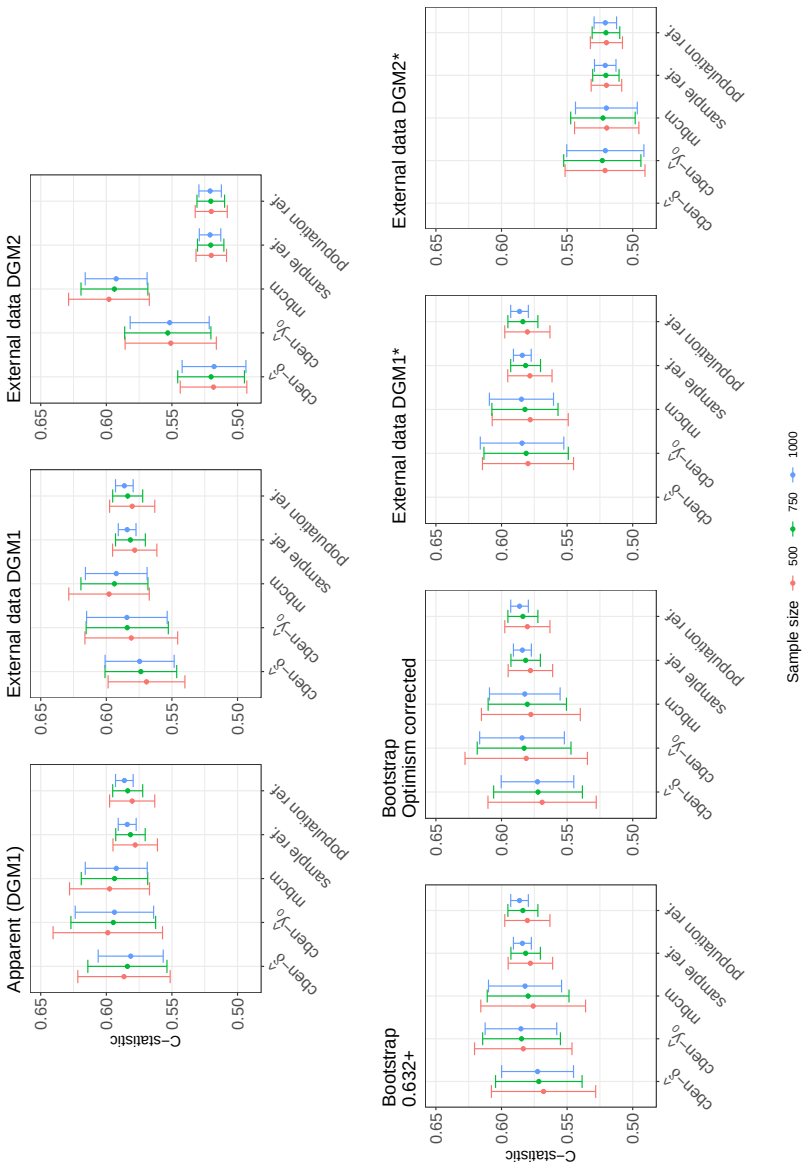

Apparent (DGM1)
External data DGM1
External data DGM2
Bootstrap 0.632+
Bootstrap Optimism corrected
External data DGM1*
External data DGM2*
C−statistic
0.65
0.60
0.55
0.50
cben−δ
cben−γ
mbcn
sample ref.
population ref.
Sample size
500
750
1000

Figure 5.1: Simulation results for the discrimination statistics in terms of mean c-statistic ± 1 SD. Top row: apparent evaluations in the original data (left), new data from the same DGM (middle), and new data from a different DGM (right). Bottom row, left to right: adjusted evaluations in the original data (bootstrap corrected 0.632+ and optimism correction), adjusted evaluations in new data from the same DGM, and adjusted evaluations in new data from a different DGM.

With respect to evaluation in new data from a different DGM (top-right), the large systematic error for cben-$\hat{y}^0$ and mbcb is apparent. For the cben-$\hat{y}^0$, this is because it relies on predictions $\hat{g}^0$ for 1:1 matching that are not suitable for the data at hand. Consequently, the observed outcome difference within matched pairs cannot be fully attributed to treatment, resulting in biased estimates. For the mbcb, this actually a feature, showing the expected model performance adjusted to the case-mix in V2 assuming the ITE model is correct. Local estimates of $\hat{P}_{\text{benefit},k,l}$ are required for actual external validation. Lastly, the original cben-$\hat{\delta}$ was a little too low (as in V1), but still quite close to the reference standards, with 1:1 matching on $\hat{\delta}$ apparently reasonable for this DGM.

Sub-figures and rows indicated with a star (in Figure 5.1 and Table 5.1) show the results after local estimation of control outcome risk (for the cben-$\hat{y}^0$) and $\hat{P}_{\text{benefit},k,l}$ (for the mbcb). For the cben-$\hat{y}^0$, this is required for accurate matching in new data.[5] For the mbcb, local estimates of $\hat{P}_{\text{benefit},k,l}$ are always required for validation purposes. For both V1 and V2, this results in essentially unbiased estimates for both the cben-$\hat{y}^0$ and the mbcb. However, rmse of the cben-$\hat{y}^0$ was still large and the mbcb is clearly to be preferred in terms of error when compared against the population reference estimates.

Summarizing, the cben-$\hat{\delta}$ showed some bias in all settings but was very stable throughout. Both the cben-$\hat{y}^0$ and mbcb were essentially unbiased when evaluated in external data, but only the latter was sufficiently precise to also outperform the cben-$\hat{\delta}$ in terms of rmse. Lastly, bootstrap procedures removed optimism across the board, but also increased variability.

[5] External data set V1 is an exception, since the DGM was exactly the same as for the development set, but this is never known in practice

5.5.3 Calibration results

Simulation results for the calibration estimates are shown in Figure 5.2 (slopes), online supplementary material Figure S5.1 (intercepts) and Table 5.2. Tabulated results corresponding to these figures is in online supplementary Tables S5.2 and S5.3. As expected, the apparent intercept and slope evaluations were uniformly 0 and 1 respectively. While this is a useful check of procedures, it also illustrates a challenge in calibration procedures: the apparent assessment is not just optimistic, but wholly uninformative.

Naive calibration assessment in V1 (*i.e.*, with offset $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ based on the ITE model under evaluation) showed optimistic slope estimates. Performing the same assessment in the whole population for DGM-1 ('population naive') gave similarly biased estimates. These findings are exaggerated when assessing calibration in V2, with the naive findings for all sample sizes seemingly very good, yet with large deviation from the true slopes for either the sample reference (*i.e.*, given the distribution of covariates in the validation data) or the population reference. Both bootstrap procedures removed optimism from apparent estimates, but the 0.632+ estimate was on average 0.062 too low for all sample sizes. Optimism correction performed better and was on average 0.036 below the population reference. Nonetheless, in terms of rmse, bootstrap estimates were all worse than the non-informative apparent evaluation. This implies that there does not seem to be enough information in a single sample to obtain reliable ITE calibration estimates.

The only consistently unbiased estimates were obtained in external data V1 and V2 after locally estimating a model for $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ in the control arm. Note that this approach does not use data twice, since the ITE calibration model is fitted in the treated arm only. Regardless of sample size or data generating mechanism, these estimates were almost unbiased and had the best rmse as compared to any of the other estimates. The only exception was for n1000 in V1, which had a similar rmse based on the original model for $\hat{g}_{lp,0}(\boldsymbol{x}_j)$. Since D and V1 are both from DGM-1, re-estimation of $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ is only beneficial if V1 has a larger control arm, which was the case for n500 and n750 for D.

Statistic	cben-$\hat{\delta}$	cben-$\hat{y}^0$	mbcb
Development data			
Apparent	0.036, 0.031, 0.025	0.045, 0.035, 0.030	**0.035, 0.027, 0.023**
0.632+	0.041, 0.035, 0.030	0.037, 0.030, 0.026	0.039, 0.030, 0.027
Opt. corrected	0.042, 0.036, 0.030	0.046, 0.036, 0.031	0.037, 0.029, 0.026
External			
DGM-1	0.028, 0.028, 0.028	0.032, 0.030, 0.030	0.036, 0.027, 0.023
DGM-2	0.023, 0.024, 0.024	0.042, 0.043, 0.041	0.085, 0.079, 0.076
DGM-1*	na	0.031, 0.030, 0.031	**0.024, 0.023, 0.024**
DGM-2*	na	0.028, 0.028, 0.029	**0.022, 0.023, 0.022**

Table 5.1: Root mean squared error against population reference as averaged over simulation runs for each measure and for each of the sample sizes (500, 750, and 1000; left to right)
*after local estimation of control outcome risk (for cben-$\hat{y}^0$) and $\hat{P}_{\text{benefit},k,l}$ (for mbcb). Bold numbers denote the best performance for each sample size in the following groups: development data, external data from DGM-1, and external data from DGM-2.

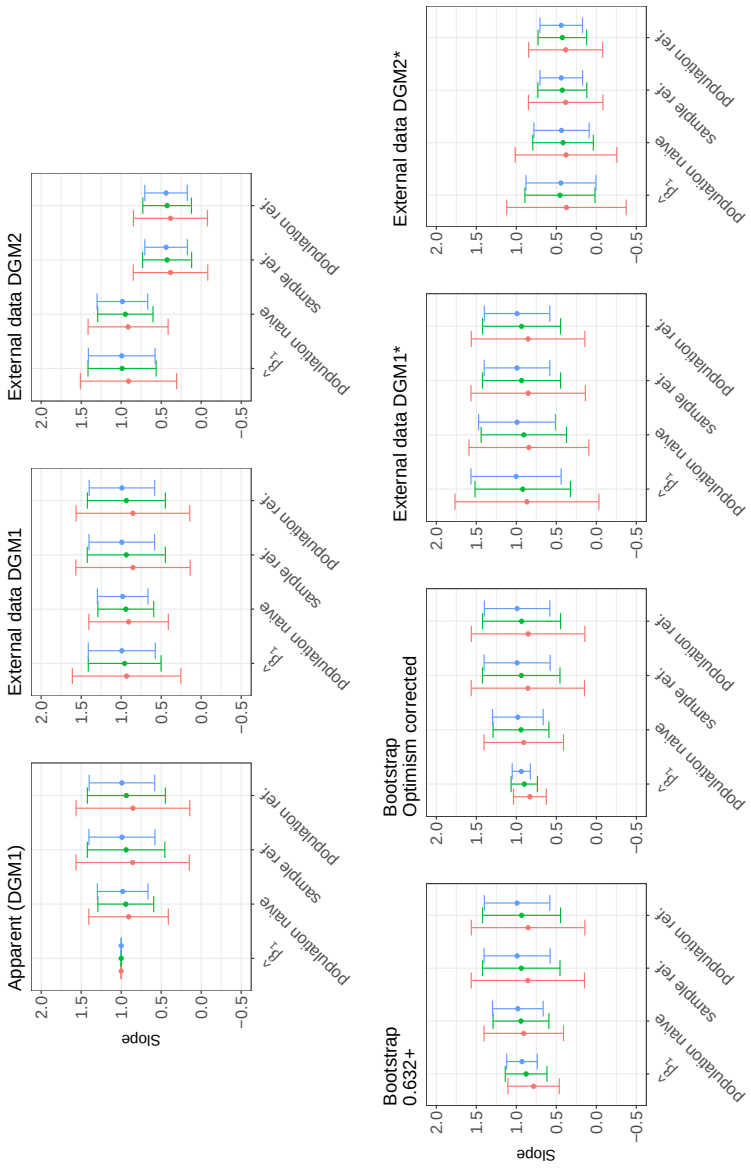
Nonetheless, since rmse is on the scale of the estimates, the absolute size of the errors was still large, casting doubt on the practical utility of ITE calibration.

Summarizing, relying on the ITE model for predictions $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ (control outcome risk) can induce large bias and local estimation of $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ in independent data is preferable. Bootstrap estimates removed optimism and performed well in terms of bias, but were highly variable for particular data sets, which limits its practical applicability. In external validation data, performance of the ITE calibration metrics provided a large improvement over apparent and bootstrap estimated in terms of both bias and root mean squared prediction error.

Statistic	$\hat{\beta}_0$	$\hat{\beta}_1$
Development data		
Apparent	**0.563, 0.378, 0.319**	**0.724, 0.491, 0.409**
0.632+	na	0.893, 0.679, 0.573
Opt. corrected	0.627, 0.450, 0.384	0.817, 0.600, 0.506
External		
DGM-1	0.367, 0.336, **0.315**	0.473, 0.413, **0.386**
DGM-2	0.924, 0.911, 0.898	0.730, 0.701, 0.665
DGM-1*	**0.337, 0.322,** 0.319	**0.412, 0.375**, 0.393
DGM-2*	**0.373, 0.284, 0.269**	**0.455, 0.316, 0.332**

Table 5.2: Root mean squared error against population reference as averaged over simulation runs for calibration intercept and slope estimates for each of the sample sizes (500, 750, and 1000; left to right). * after local estimation of control outcome risk. Bold numbers denote the best performance for each sample size in the following groups: development data, external data from DGM-1, and external data from DGM-2.

Figure 5.2: Simulation results for the ITE calibration slope estimates (mean ± 1 SD). Top row: apparent evaluations in the original data (left), new data from the same DGM (middle), and new data from a different DGM (right). Bottom row, left to right: adjusted evaluations in the original data (bootstrap corrected 0.632+ and optimism correction), adjusted evaluations in new data from the same DGM, and adjusted evaluations in new data from a different DGM.


Apparent (DGM1)
External data DGM1
External data DGM2
Bootstrap 0.632+
Bootstrap Optimism corrected
External data DGM1*
External data DGM2*
Slope
2.0
1.5
1.0
0.5
0.0
−0.5
β_1
population naive
sample ref.
population ref.
Sample size
500
750
1000

5.6 Applied example: the third International Stroke Trial

Patients with an ischemic stroke have sudden onset of neurological symptoms due to a blood clot that narrows or blocks an artery that supplies the brain. A key component in the emergency medical treatment of these patients includes clot-busting drug alteplase (intravenous thrombolysis recombinant tissue-type plasminogen activator) [161].

The third International Stroke Trial (IST-3) was a randomized trial and investigated the benefits and harms of intravenous thrombolysis with alteplase in acute ischemic stroke [148]. This large trial included 3035 patients receiving either alteplase or placebo in a 1:1 ratio. The primary outcome was proportion of patients that was alive and independent at 6-month follow-up, which we used as outcome of interest here. Primary analyses of the treatment effect were performed by with logistic regression adjusted for linear effects of age, National Institutes of Health stroke scale (NIHSS) score, time from onset of stroke symptoms to randomization, and presence (vs absence) of ischemic change on the pre-randomization brain scan according to expert assessment. This analysis showed weak evidence of an effect (OR 1.13, 95% CI 0.95-1.35), but subgroup analyses suggested possibly heterogeneous treatment effect by age, NIHSS score, and predicted probability of a poor outcome.

For illustrative purposes, we here compare a main effects logistic regression model similar to the original adjusted analysis (model 1) with a model where all covariate-treatment interactions were included (model 2). The outcome was coded as 0 for those independent and alive after 6 months and 1 otherwise. The included variables were treatment, age, NIHSS, time (from onset of stroke symptoms to randomization), and imaging status (presence vs absence of ischemic change on the pre-randomization brain scan). Continuous variables age, NIHSS, and time, were modeled using smoothing splines. We also included covariate-treatment interactions for these variables. Continuous variables age, NIHSS, and time, were modeled using smoothing splines with shrinkage. We also included covariate-treatment interactions for these variables. Models were fitted using the `mgcv` package in `R` with defaults smoothing parameter selection based on generalized cross-validation [83]. All in all, this applied example illustrates different ways to assess the quality of individualized treatment effect predictions. The evaluated models were emphatically chosen for this purpose and were not developed in collaboration with clinical experts in the field. Hence,

they are not meant to me applied in practice.

The exact parameter estimates for both models are not of key interest, but the apparent performance with respect to outcome risk prediction was good for both: c-statistics were 0.826 and 0.831 for model 1 and 2 respectively, with accompanying Brier scores of 0.160 and 0.158 and Nagelkerke R^2 of 0.389 and 0.402. The Spearman correlation between outcome risk predictions for both models, conditional on the assigned treatments, was 0.99.

Model	cben-$\hat{\delta}$	cben-$\hat{y}^0$	mbcb	$\hat{\beta}_0$	$\hat{\beta}_1$
Apparent					
M1	0.488	0.489	0.510	-0.117	
M2	0.562	0.570	0.567	0.011	1.071
bootstrap 0.632+					
M1	0.489	0.499	0.505		
M2	0.535	0.559	0.536		0.522
Optimism corrected					
M1	0.485	0.475	0.507	-0.068	
M2	0.534	0.544	0.518	-0.022	0.895

Table 5.3: Applied example discrimination and calibration statistics for predicted individualized treatment effect.

Nonetheless, the range of predicted ITEs (*i.e.*, on the risk difference scale) was very different. Model 1 predicted ITEs with median -0.020 (IQR -0.027, -0.011 and range -0.029, 0.000), while model 2 predicted ITEs with median -0.026 (IQR -0.059, 0.027 and range -0.811, 0.213). That is, the predicted treatment effect was very similar across individuals when predicted by model 1 (assuming a constant treatment effect on the log odds scale), but not when predicted by model 2 (assuming a heterogeneous treatment effect on the log odds scale). Table 5.3 shows the apparent and bootstrap corrected results for discrimination and calibration assessment at the ITE level for the applied example as averaged over 1000 bootstrap samples.

With respect to ITE discrimination, both apparent and bootstrap-corrected discrimination estimates favored model 2 over model 1, with model 1 estimates around the no discriminative ability value of 0.5. If model 2 were entirely correct, the expected c for benefit for samples with similar characteristics was estimated

to be 0.567 (apparent mbcb). While we know that model 2 is just a model and not exactly correct, this value is relevant since it provides the upper bound of ITE concordance for the combination of model and data. Subsequently, the bootstrap procedures uncovered evidence of overfitting of ITE's and provided downward adjusted estimates.

Calibration slope estimates suggested that model 2 ITE estimates are more heterogeneous than justified by the data, and require shrinkage. The amount of shrinkage suggested varies considerably between the 0.632+ and optimism corrected estimates. Based on the simulation study, optimism correction was already conservative and was to be preferred over the 0.632+ slope which were yet more conservative. Note that the calibration slope for model 1 is not estimable (since the ITEs have no variability on the logit scale) and the intercept estimate for model 1 clearly show that the degree of predicted benefit is underestimated.

In summary, the results indicate that model 1 did not provide useful differentiation in terms of ITEs. While the discriminative ability of model 2 seems modest, clear benchmarks are lacking. After updating based on the optimism corrected calibration estimates, model 2 ITE predictions may still be meaningful, having a median of -0.02 (IQR -0.52, 0.02 and range -0.58, 0.19). Comparing the 1969 patients predicted to have benefit ($\hat{\delta}_{model2} < 0$) with the remaining 1066 patients ($\hat{\delta}_{model2} \geq 0$), the first were older [median(IQR) age 83 (78-87) vs 73 (63-82)], had worse symptoms [median(IQR) nihss 15(10-20) vs 6(4-9)], were treated earlier [median(IQR) time in hours 3.5 (2.5-4.9) vs 4.2 (3.6-4.8)], were more likely to have visual infarction on imaging (43% vs 36%), and had a less favorable outcome on average (alive and independent after 6 months in 23% vs 58% respectively).

5.7 Software

`R` package **iteval** (`https://github.com/jeroenhoogland/iteval`) provides a free software implementation on the freely available `R` software environment for statistical computing [160] for the cben-$\hat{\delta}$, cben-$\hat{y}^0$, mbcb, and calibration measures as defined in this paper.

5.8 Discussion

Measures of calibration and discrimination have a long history in the context of prediction models for observed outcome data, especially of the binary type. However, the evaluation of individualized treatment effect (ITE) prediction models is more challenging, first and foremost because of the causal nature of the predictions and the ensuing unobservable nature of individualized treatment effects. In this paper, we utilized the potential outcomes framework [93] to obtain insight into existing performance measures [132] and to develop novel measures of discrimination and calibration for ITE prediction models. We proposed model-based statistics to address challenges of existing methods. Importantly, these statistics are applicable regardless of the modeling approach used to generate ITE predictions, as long as predictions for each potential outcome are available. This means that our methods are usable for both statistical, as well as machine learning methods. Also, while the primary focus was on dichotomous outcomes, we also provided residual-based approaches for continuous outcome models. As such, our work provides generally applicable tools for the endeavor of ITE prediction model evaluation [146].

With respect to discriminative ability, the model-based c-for-benefit (mbcb) provides both a normal performance measure and an expected (case-mix adjusted) reference level for new data. The latter is relevant since concordance probabilities are known to be sensitive to case-mix [159]. Also, bootstrap procedures are available to adjust for optimism during model development. In the simulation study, the mbcb estimates were best in terms of both bias and root mean squared error across simulation settings. Both matching-based measures of discriminative performance had specific downsides. The original cben-$\hat{\delta}$ has a difficult interpretation and was downward biased in the simulation study, but was very stable throughout. The adaptations implemented in the cben-$\hat{y}^0$ did remove the bias, but at the cost of much larger variability. We hypothesize that the stability of the mbcb is due to the lack of a need for a matching algorithm. The large variability of cben-$\hat{y}^0$ likely relates to strong reliance of the matching procedure on the accuracy of predicted control outcome risk.

With respect to calibration, the potential outcomes framework provided a model-based method that evaluates ITE prediction in the treated individuals, against an offset of prediction outcome risk under control treatment. Compared to traditional calibration measures at the level of the outcome of interest, calibration of ITE predictions has the additional challenge that ITEs are in fact relative effects. As such, correct calibration of ITEs depends on correct calibration of

outcome risk under control treatment. In line, local updating of the model predicting control outcome risk proved paramount for valid assessment of ITE calibration.

A key finding for both ITE discrimination and calibration measures was that bootstrap procedures were able to remove optimism (*i.e.* reduce bias), but that the increase in variance of the estimator generally led to increased root mean squared error. This implies that external data is required to accurately assess ITE predictions. The underlying reason is the need for local estimates (*i.e.*, independent of the ITE model under evaluation) of control and treated outcome risk. While these steps were incorporated in the bootstrap procedures, they are necessarily noisy since they have to rely on only 36,8% of the data for any particular bootstrap run.

While this paper focused on measures specifically targeting ITE predictions, in practice we recommend assessing prediction performance with respect to the observed outcomes first [8, 9]. For instance, performance with respect to outcome risk can be evaluated in the control arm and in the treated arm separately. If performance is good, one can move on to ITE evaluation. The motivation for this hierarchy is that ITEs reflect differences and that they hence compound errors in both potential outcome predictions.

Limitations of the current work include the limited nature of the simulation study which was mainly performed for illustrative purposes. While both discrimination and calibration are well researched in classical settings, their application to ITE predictions is relatively novel. While we did elucidate several aspects of ITE prediction model evaluation in terms of discrimination and calibration, important questions remain. These include questions with respect to the best strategy for model comparison, the uncertainty of the estimated statistics, the relation between discrimination and calibration on the outcome risk level and the ITE level, and the relation between discrimination and calibration statistics and clinical usefulness of the models. With respect to uncertainty estimates, bootstrap procedures provide a good option, but many of the challenges are still open.

In terms of future research, it would be interesting to evaluate whether some level of grouping is beneficial for the evaluation of model performance. Paradoxically, the aim for precision underlying the development of ITE models may hamper the possibility to evaluate them, since individual level treatment effects are inherently unobservable, and their evaluation hence involves approximations based on the very model under evaluation. Also, especially in the binary event

case, even if individual-level treatment effects would be observable, they would still be very noisy. This is the underlying reason that the c-statistics for benefit are so much lower than c-statistics on the outcome level.

Summarizing, we used the potential outcomes framework to obtain insight into measures of discrimination and calibration at the level of individualized treatment effect predictions. This allowed for a principled examination of existing measures and a proposal of new measures that may enrich the toolbox for ITE model evaluation. Further research is necessary to improve understanding of the exact characteristics of these measures under varying conditions with respect to sample size, degree of treatment effect heterogeneity, and explained variation.

Acknowledgements

This project received funding from the European Union's Horizon 2020 research and innovation program under ReCoDID grant agreement No 825746. Jeroen Hoogland and Thomas P. A. Debray acknowledge financial support from the Netherlands Organisation for Health Research and Development (grant 91215058). Thomas P. A. Debray also acknowledges financial support from the Netherlands Organisation for Health Research and Development (grant 91617050). Orestis Efthimiou was supported by the Swiss National Science Foundation (Ambizione grant number 180083). We like to thank the researchers involved in the original stroke trial for use of their data [148, 162].

Data Availability

Data for the International Stroke Trial-3 applied example are publicly available [162]. `R` package **iteval** is available on GitHub (`https://github.com/jeroenhoogland/iteval`) and provides functions to derive the cben-$\hat{\delta}$, cben-$\hat{y}^0$, mbcb, and calibration measures as defined in this paper. Github repository **iteval-sims** (`https://github.com/jeroenhoogland/iteval-sims`) provides the required files and instructions for replication of the simulation study.

Supplementary Material

S5.1 Binomial outcome data

S5.1.1 Absolute risk, risk difference, and binomial error

Focusing on binary outcomes, assume we observe outcome $Y_i \in \{0, 1\}$ and covariate status $\boldsymbol{x}_i$ for each individual i. Using data on n individuals, we can model the outcome risk $P(Y_i = 1|A = a_i, \boldsymbol{X} = \boldsymbol{x}_i)$. There are two sources of error when using such a model to predict binary outcomes. There is the reducible error in modeling the risk (*i.e.*, how well the modelled probability approximates the actual probability of an event), and there is the irreducible error in the difference between the actual probability of an event and its manifestation as a $\{0, 1\}$ outcome (binomial error).

Adding to this, actual interest is in the difference in outcome risk under different treatment assignment $a \in \{0, 1\}$. That is, interest is in $p(Y_i = 1|A = 1, \boldsymbol{X} = \boldsymbol{x}_i) - p(Y_i = 1|A = 0, \boldsymbol{X} = \boldsymbol{x}_i)$. The range of possible true (and estimated) treatment effects (risk differences) includes all values in the $[-1, 1]$ interval, but the observed difference between any two outcomes can only be one of $\{-1, 0, 1\}$. An example may be helpful to appreciate the large influence of irreducible error in this setting. For instance, regardless of any modeling, assume that an active treatment (as compared to a control condition) reduces outcome risk from 25% to 20% for a certain individual. Moreover, assume that these probabilities are known exactly and that this individual can be observed under both treatment conditions. A simple probabilistic exercise[6] shows that the different outcome probabilities are $P(Y^0 = 0, Y^1 = 0) = 0.6, P(Y^0 = 0, Y^1 = 1) = 0.15, P(Y^0 = 1, Y^1 = 0) = 0.2$, and $P(Y^0 = 1, Y^1 = 1) = 0.05$. That is, the probability that the active treatment induces any observed outcome difference is 35%, and only 20% is in the expected direction (*i.e.* in the direction of the treatment effect). This is just due to the *irreducible* error, apart from any modeling issues, and ignoring the fact that in practice only one potential outcome is observed of each individual. The insensitivity of binary endpoints is of course well known in the context of trials, where a larger number of replications can provide a solution when the average treatment effect is of interest. In the case of individualized treatment effect estimation however, the required number of replications is more challenging to control due to its complex dependence on all individual-level

[6] For instance, $P(Y^0 = 0, Y^1 = 0) = (1 - P(Y^0 = 0))(1 - P(Y^1 = 0)) = (1 - 0.25)(1 - 0.2) = 0.6$

characteristics of interest.

S5.1.2 Scale matters

Models that predict the risk of a binary event commonly make use of a link function in order to map a function of the covariates in $\mathbb{R}$ onto the probability scale [163]. Such link functions, such as the logit or inverse Gaussian, are inherently non-linear and hence do not preserve additivity. Consequently, a treatment effect that is constant (*i.e.*, does not vary with other covariates) before applying the link function *shall* vary with other covariates on the risk scale and vice versa. As an example, we write h^{-1} for an inverse link function and take control risk to be a function $f(\cdot)$ of only one random variable X (*i.e.*, $P(Y^{a=0}|X = x) = h^{-1}(f(X))$). Subsequently, assume a constant (homogeneous) relative treatment effect d such that $P(Y^{a=1}|X = x) = h^{-1}(f(X) + d)$, then the absolute treatment effect necessarily depends on X, since

$$\begin{aligned}\delta(x) &= P(Y^{a=1}|X = x) - P(Y^{a=0}|X = x) \\ &= h^{-1}[f(X) + d] - h^{-1}[f(X)] \neq h^{-1}[f(X) + d - f(X)] = h^{-1}(d)\end{aligned} \tag{S5.1.1}$$

unless $h^{-1}(\cdot)$ is linear. Consequently, between-individual variability (*i.e.*, variability in terms of X) directly changes control outcome risk *and* affects the absolute effect of d on the probability scale even if d is constant. For instance, a constant treatment effect on the log-odds scale translates into heterogeneous treatment effect on the risk difference scale. Thereby, relatively simple treatment effect structures may lead to meaningful between-individual treatment effect variability at the risk difference level if there is large variability in $h^{-1}[f(X)]$ [146, 164]. In addition, treatment effect may interact with X in the domain of $h^{-1}(\cdot)$, *i.e.*, we may directly model treatment effect heterogeneity. These two sources of variability in $\delta(x)$ can no longer be discerned when evaluating just the estimates $\hat{\delta}(x)$. Hence, the benefit in terms of interpretation of measures on the scale of $\delta(x)$ [92], as of interest in this paper, has a price in that they conflate variability in $\hat{\delta}(x)$ from different sources: between-subject variability in $P(Y^{a=0}|X = x)$ and genuine treatment effect heterogeneity on the scale used for modeling.

S5.2 Discrimination estimand

Harrell's c-statistic [149, 8] can be applied to ordered predictions and (possibly censored) ordered outcomes. Applying Harrell's c-statistic with ITE predictions

and within-individual differences in potential outcomes $Y_k^{a=1} - Y_k^{a=0}$ as simulated from some data generating mechanism, the equation for the concordance probability can be written as

$$c_{\hat{\delta},ben} = \frac{\sum_k \sum_{l \neq k} \left[I(\hat{\delta}_k < \hat{\delta}_l)\text{ben}_{kl} + \frac{1}{2} I(\hat{\delta}_k = \hat{\delta}_l)\text{ben}_{kl} \right]}{\sum_k \sum_{l \neq k} [\text{ben}_{kl}]} \tag{S5.2.1}$$

with

$$\text{ben}_{kl} = I([Y_k^{a=1} - Y_k^{a=0}] < [Y_l^{a=1} - Y_l^{a=0}]) \tag{S5.2.2}$$

However, the within-individual differences in sampled potential outcomes $Y_k^{a=1} - Y_k^{a=0}$ are just a single manifestation of treatment effect for covariate matrix $\boldsymbol{X}$, and interest is in the expected value over repeated samples of potential outcomes given $\boldsymbol{X}$. Taking this expectation $\mathbb{E}_{\boldsymbol{Y}^{a=A}|\boldsymbol{X}}$ over equation (S5.2.1) does not affect ITE predictions, since these are invariant conditional on a fixed ITE model and fixed $\boldsymbol{X}$. For ben_{kl},

$$\begin{aligned} \mathbb{E}_{\boldsymbol{Y}^{a=A}|\boldsymbol{X}}(\text{ben}_{kl}) &= \mathbb{E}_{\boldsymbol{Y}^{a=A}|\boldsymbol{X}}(I([Y_k^{a=1} - Y_k^{a=0}] < [Y_l^{a=1} - Y_l^{a=0}])) \\ &= P([Y_k^{a=1} - Y_k^{a=0}] < [Y_l^{a=1} - Y_l^{a=0}]) \\ &= P_{\text{benefit},k,l} \quad \text{(as defined in Section 5.3.4)} \end{aligned} \tag{S5.2.3}$$

In turn, substituting an estimate of $P_{\text{benefit},k,l}$ for ben_{kl} in equation (S5.2.1) gives the equation for the mbcb (equation (5.3.10)). Instead substituting the true probabilities $P_{\text{benefit},k,l}$, as known in a simulation context given the data generating mechanism, provides the expected value for the ITE concordance statistic for a given data generating mechanism, a fixed matrix of observed covariate values $\boldsymbol{X}$, and a fixed ITE model. Therefore, it was used as the estimand value of the ITE concordance statistic for a given sample (denoted 'sample reference').

When the sample size is very large, such as for the simulated populations of size 100,000, calculation of the mbcb is computationally very intensive and an approximation based on $c_{\hat{\delta},ben}$ (equation (S5.2.1)) is accurate. Note that $c_{\hat{\delta},ben}$ is still unbiased, but of course more variable.

S5.3 Performance evaluation details

S5.3.1 Apparent performance

Apparent ITE model performance was evaluated, without any adjustment, in the development sample D in which the ITE model was fitted. Consequently, apparent estimates can be expected to be optimistic. For apparent calibration performance of a logistic ITE model based on maximum likelihood estimation, note that the estimates will invariably be $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ since the calibration model is of the exact same type. All in all, apparent performance was primarily assessed to show the need for internal validation procedures that correct for optimism or, better yet, new data.

S5.3.2 Internal validation

Discrimination
Internal validation was performed based on a non-parametric bootstrapping procedure based on 100 bootstrap samples. Performance estimates were based on either a 0.632+ method [165] adapted for application in the context of c-statistics or on optimism correction [8].

The adapted 0.632+ method provides a weighted average of apparent performance and average out-of-sample performance as based on predictions from bootstrap models for the cases not in the bootstrap sample. Writing $\hat{c}_{app}$ (scalar) for the apparent c-statistic and $\hat{c}_{oos}$ (scalar) for the average out-of-sample c-statistic across bootstrap replications,

$$\hat{c}_{oos} = \begin{cases} \min(\gamma, \hat{c}_{oos}), & \hat{c}_{app} \geq \gamma \\ \max(\gamma, \hat{c}_{oos}), & \hat{c}_{app} < \gamma \end{cases} \tag{S5.3.1}$$

$$R = \begin{cases} \frac{|\hat{c}_{app} - \hat{c}_{oos}|}{|\hat{c}_{app} - \gamma|}, & |\hat{c}_{oos} - \gamma| < |\hat{c}_{app} - \gamma| \\ 0, \quad \text{otherwise} \end{cases} \tag{S5.3.2}$$

$$w = \frac{0.632}{1 - 0.368R} \tag{S5.3.3}$$

$$\hat{c}_{0.632+} = \hat{c}_{app}(1 - w) + w\hat{c}_{oos} \tag{S5.3.4}$$

where γ is the value of the statistic for an uninformative model (so $\gamma = 0.5$ for c-statistics), and w is a weight that depends on the discrepancy between

apparent and out-of-sample performance. To prevent that R falls outside of the $(0, 1)$, we avoid the possibility of bootstrap correction towards a point beyond the no information threshold by replacement of $\hat{c}_{oos}$ with $\hat{c}'_{oos}$ throughout, with

$$\hat{c}'_{oos} = \begin{cases} \min(\gamma, \hat{c}_{oos}), & \hat{c}_{app} \geq \gamma \\ \max(\gamma, \hat{c}_{oos}), & \hat{c}_{app} < \gamma \end{cases} \tag{S5.3.5}$$

Subsequently, R reflect the degree of overfitting and ranges from zero to one, with w depending only on R and ranging from 0.632 and 1. Thereby, $\hat{c}_{0.632+}$ moves towards $\hat{c}'_{oos}$ when the amount of overfitting ($|\hat{c}_{app} - \hat{c}'_{oos}|$) is large with respect to the models gain relative to no information ($|\hat{c}_{app} - \gamma|$). The choice to use $\hat{c}'_{oos}$ instead of $\hat{c}'_{oos}$ in (S5.3.5) was to avoid correction of an apparent estimate beyond the no information threshold.

Alternatively, optimism correction estimates optimism as the average difference between performance of bootstrap models as evaluated in a) the original full data set D and b) within the bootstrap sample. In case of overfitting, the discrepancy between the two will increase. The apparent estimate is subsequently corrected for this bootstrap estimate of optimism.

Obtaining either the 0.632+ or optimism corrected estimates for cben-$\hat{\delta}$ and cben-$\hat{y}^0$ is straightforward. One subtlety is that in case of unequal group sizes (treated vs control), the average over 1000 repeated analyses of subsamples of the larger arm was taken to accommodate for 1:1 matching. For the model-based estimates, a choice with respect to the estimation of $\hat{P}_{\text{benefit},k,l}$ has to be made with respect to out-of-sample evaluation. To avoid bias, the out-of-sample evaluation of $\hat{P}_{\text{benefit},k,l}$ for the 0.632+ estimate was based on a model for $\hat{g}_0$ and $\hat{g}_1$ is the out-of-sample cases (with the same specification as the model under evaluation). For the optimism correction, $\hat{P}_{\text{benefit},k,l}$ for the whole of D was based on the ITE model as developed in the full sample D. That is, the 0.632+ model-based c-statistic estimates were obtained from (1) out-of-sample predictions $\hat{\delta}(\boldsymbol{x}_{i \in oos})$ from bootstrap models and (2) $\hat{P}_{\text{benefit},k,l}$ based on an out-of-sample model. Optimism corrected model-based c-statistic estimates were obtained from (1) predictions $\hat{\delta}(\boldsymbol{x}_i)$ from bootstrap models and $\hat{P}_{\text{benefit},k,l}$ based on the development model.

Calibration
Bootstrap evaluation of the calibration parameters was also performed. A

0.632+ estimate was derived for the slope estimates in analogy to the derivation for c-statistics, but using $\gamma = 0$ for the value that slope β_1 takes for an uninformative model. Out-of-sample estimates of $\hat{g}_0(\cdot)$ were based on a model fitted in just the out-of-sample controls to serve as an offset in the calibration model fitted in the out-of-sample treated arm. A 0.632+ estimate for the calibration intercept parameter is not readily available since a γ value for a non-informative intercept cannot be defined. Optimism corrected bootstrap estimates were obtained for both intercepts and slopes. Estimates of $\hat{g}_0(\cdot)$ in the bootstrap sample were based on the bootstrap model and estimates $\hat{g}_0(\cdot)$ in the original data were based on the original ITE model fitted in development data D.

S5.3.3 External validation

Discrimination
External validation was performed in both V1 (DGM-1) and V2 (DGM-2). ITE predictions can be evaluated directly using cben-$\hat{\delta}$. For cben-$\hat{y}^0$, which matches based on predicted control outcome risk $\hat{g}_0(\cdot)$, a key question is whether $\hat{g}_0(\cdot)$ may best be based on the ITE model *or* on a *new* model fitted in the control arm of the external data. In practice, the accuracy of $\hat{g}_0(\cdot)$ based on the ITE model can be assessed in the control arm of the external data. If not satisfactory, a new model for $\hat{g}_0(\cdot)$ can be derived in the external data for use with the cben-$\hat{y}^0$. The latter option was taken for the simulation study based on a refitting of the relevant parts of model (5.5.3) in the control arm (*i.e.*, omitting parameters relating to a which equal 0 for controls) of the external data. Note that fitting a new model will in general remove bias, but may have a high cost in terms of variance if the external data set is small. For the model-based c-for-benefit (mbcb), the accuracy of $\hat{P}_{\text{benefit},k,l}$, and hence the underlying $\hat{g}_0(\cdot)$ and $\hat{g}_1(\cdot)$, is paramount. In the mbcb, $\hat{P}_{\text{benefit},k,l}$ is the sole carrier of information from the external data and acts as the reference for ITE predictions under evaluation. In line with the procedure for the cben-$\hat{y}^0$, performance with respect $\hat{g}_0(\cdot)$ and $\hat{g}_1(\cdot)$ can be examined in the external data (control arm and treated arm respectively) and may indicate the need for a new model. Again, the latter option was chosen for the simulation study. Note that D, V1 and V2 were always of equal size, such that there was the benefit of possibly reducing bias while avoiding the possible harm of increased variance. Finally, note that while cben-$\hat{\delta}$, cben-$\hat{y}^0$, and the mbcb focus on $\hat{\delta}(\boldsymbol{x}_i)$, inadequate prediction performance with respect to $\hat{g}_0(\cdot)$ and/or $\hat{g}_1(\cdot)$ is an ominous sign for ITE model performance, and its evaluation should in practice be part of any performance evaluation with respect to ITEs.

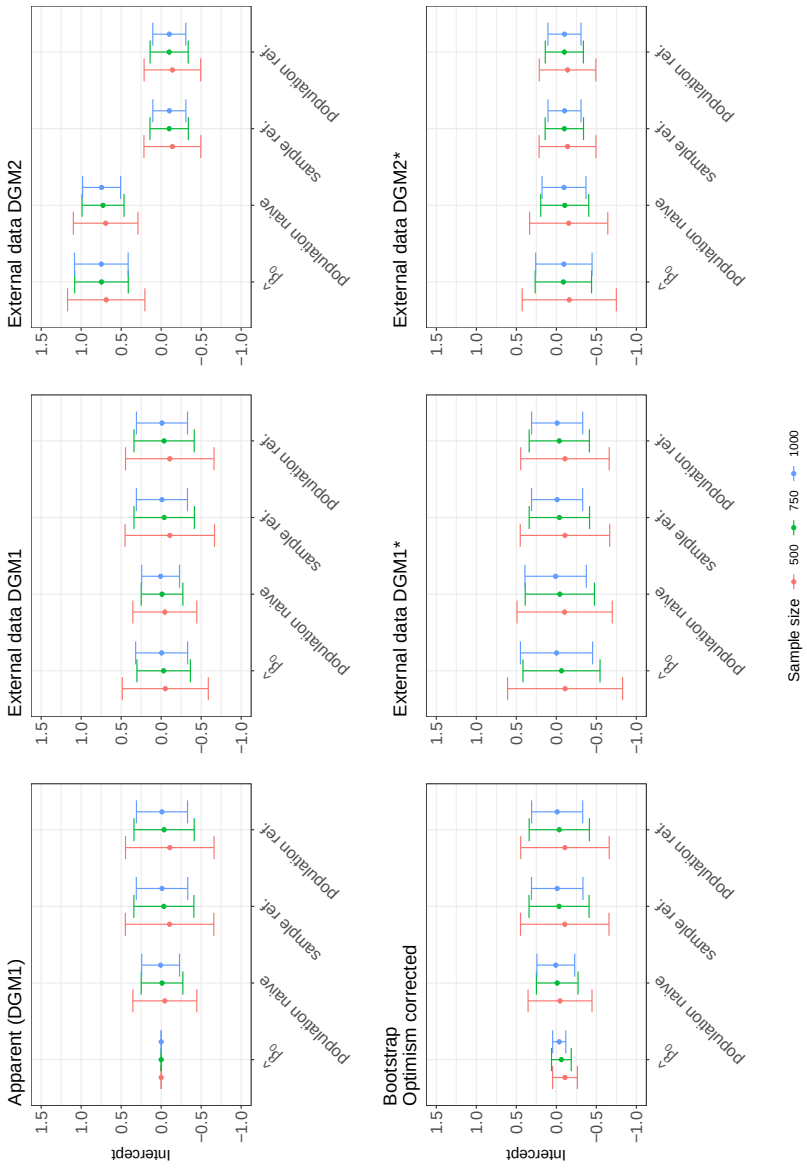
Calibration
Direct calibration assessment in external data exactly followed the lines of apparent calibration assessment with all predictions (both $\hat{\delta}_{lp}(\boldsymbol{x}_j)$ and $\hat{g}_{lp,0}(\boldsymbol{x}_j)$) based on the ITE model as derived in D and applied in V1 and V2. Adjusted estimates were obtained based on local predictions $\hat{g}_{lp,0}(\boldsymbol{x}_j)$ after refitting of the relevant parts of model (5.5.3) in the control arm (*i.e.*, omitting parameters relating to a which equal 0 for controls) of the external data.

As for discrimination, the estimands as defined in Section 5.5 reflect the target reference parameter values for a specific ITE model as evaluated in a specific sample (D, V1 or V2). To remove dependence of the performance measure on a (small) sample, population reference values were derived per data generating mechanism as for the discrimination measures. Derivation was exactly analogous to the description under 'estimands' in Section 5.5. In addition, a naive population reference was derived for calibration assessment that mirrored assessment in D (*i.e.*, naive referring to adjustment based on $\hat{g}_0(\cdot)$ based on the ITE model instead of independent data). Therein, this naive population reference helps to remove the influence of sample size from the assessment of bias due to misspecification of $\hat{g}_0(\cdot)$.

S5.4 Additional simulation study results

(Continues on the next page)

Figure S5.1: Simulation results for the ITE calibration intercept estimates (mean ± 1 SD). Top row: apparent evaluations in the original data (left), new data from the same DGM (middle), and new data from a different DGM (right). Bottom row, left to right: adjusted evaluations in the original data (bootstrap corrected 0.632+ and optimism correction), adjusted evaluations in new data from the same DGM (middle), and adjusted evaluations in new data from a different DGM (right)


Apparent (DGM1)
Bootstrap
Optimism corrected
External data DGM1
External data DGM1*
External data DGM2
External data DGM2*
Intercept
1.5
1.0
0.5
0.0
−0.5
−1.0
$\hat{\beta}_0$
population naive
sample ref.
population ref.
Sample size
500
750
1000

Statistic	cben-$\hat{\delta}$	cben-$\hat{y}^0$	mbcb	sample ref.	population ref.
n=500					
Development data					
Apparent	0.587	0.599	0.598	0.578	0.580
0.632+	0.568	0.583	0.576	0.578	0.580
Opt. corrected	0.569	0.581	0.578	0.578	0.580
External, n=500					
DGM-1	0.569	0.581	0.598	0.578	0.580
DGM-2	0.518	0.551	0.598	0.520	0.520
DGM-1*		0.580	0.578	0.578	0.580
DGM-2*		0.521	0.520	0.520	0.520
n=750					
Development data					
Apparent	0.584	0.595	0.594	0.582	0.584
0.632+	0.572	0.585	0.580	0.582	0.584
Opt. corrected	0.572	0.583	0.580	0.582	0.584
External					
DGM-1	0.574	0.584	0.594	0.582	0.584
DGM-2	0.520	0.553	0.594	0.520	0.520
DGM-1*		0.581	0.582	0.582	0.584
DGM-2*		0.523	0.523	0.520	0.520
n=1000					
Development data					
Apparent	0.581	0.594	0.592	0.584	0.586
0.632+	0.573	0.585	0.582	0.584	0.586
Opt. corrected	0.573	0.584	0.582	0.584	0.586
External					
DGM-1	0.575	0.584	0.592	0.584	0.586
DGM-2	0.518	0.552	0.592	0.521	0.521
DGM-1*		0.584	0.585	0.584	0.586
DGM-2*		0.521	0.520	0.521	0.521

Table S5.1: Mean over simulation runs for each discrimination measure and for each of the sample sizes (500, 750, and 1000; left to right). *) After local estimation of control outcome risk (for cben-$\hat{y}^0$) and $\hat{P}_{\text{benefit},k,l}$ (for mbcb).

Statistic	$\hat{\beta}_0$	population naive	sample ref.	population ref.
Development data, n=500				
Apparent	0.000	-0.046	-0.105	-0.107
0.632+		-0.046	-0.105	-0.107
Opt. corrected	-0.108	-0.046	-0.105	-0.107
External, n=500				
DGM-1	-0.052	-0.046	-0.108	-0.107
DGM-2	0.687	0.694	-0.140	-0.140
DGM-1*	-0.109	-0.103	-0.108	-0.107
DGM-2*	-0.162	-0.154	-0.140	-0.140
Development data, n=750				
Apparent	0.000	-0.011	-0.034	-0.037
0.632+		-0.011	-0.034	-0.037
Opt. corrected	-0.063	-0.011	-0.034	-0.037
External				
DGM-1	-0.031	-0.011	-0.038	-0.037
DGM-2	0.746	0.727	-0.100	-0.100
DGM-1*	-0.064	-0.043	-0.038	-0.037
DGM-2*	-0.088	-0.105	-0.100	-0.100
Development data, n=1000				
Apparent	0.000	0.007	-0.011	-0.010
0.632+		0.007	-0.011	-0.010
Opt. corrected	-0.036	0.007	-0.011	-0.010
External				
DGM-1	-0.006	0.007	-0.009	-0.010
DGM-2	0.748	0.745	-0.102	-0.101
DGM-1*	-0.003	0.009	-0.009	-0.010
DGM-2*	-0.094	-0.096	-0.102	-0.101

Table S5.2: Mean over simulation runs for calibration intercept estimates each of the sample sizes (500, 750, and 1000; left to right). *) after local estimation of control outcome risk.

Statistic	$\hat{\beta}_1$	population naive	sample ref.	population ref.
Development data, n=500				
Apparent	1.000	0.907	0.856	0.853
0.632+	0.785	0.907	0.856	0.853
Opt. corrected	0.831	0.907	0.856	0.853
External, n=500				
DGM-1	0.932	0.907	0.852	0.853
DGM-2	0.908	0.913	0.384	0.383
DGM-1*	0.868	0.843	0.852	0.853
DGM-2*	0.374	0.379	0.384	0.383
Development data, n=750				
Apparent	1.000	0.942	0.939	0.935
0.632+	0.878	0.942	0.939	0.935
Opt. corrected	0.901	0.942	0.939	0.935
External				
DGM-1	0.956	0.942	0.935	0.935
DGM-2	0.989	0.948	0.426	0.425
DGM-1*	0.920	0.907	0.935	0.935
DGM-2*	0.455	0.416	0.426	0.425
Development data, n=1000				
Apparent	1.000	0.981	0.990	0.991
0.632+	0.930	0.981	0.990	0.991
Opt. corrected	0.939	0.981	0.990	0.991
External				
DGM-1	0.992	0.981	0.992	0.991
DGM-2	0.993	0.985	0.439	0.439
DGM-1*	1.004	0.991	0.992	0.991
DGM-2*	0.443	0.435	0.439	0.439

Table S5.3: Mean over simulation runs for calibration slope estimates each of the sample sizes (500, 750, and 1000; left to right). *) after local estimation of control outcome risk.

Chapter 6

Prognosis and prediction of antibiotic benefit in adults with clinically diagnosed acute rhinosinusitis: an individual participant data meta-analysis

Hoogland J, Takada T, Smeden M, Rovers MM, de Sutter AI, Merenstein D, van Essen GA, Kaiser L, Liira H, Little P, Bucher HC, Moons KGM, Reitsma JB, Venekamp RP. Prognosis and prediction of antibiotic benefit in adults with clinically diagnosed acute rhinosinusitis: an individual participant data meta-analysis. (*under revision*)

Abstract
A previous individual participant data meta-analysis (IPD-MA) of antibiotics for adults with clinically diagnosed acute rhinosinusitis (ARS) showed average effectiveness of antibiotics, but was unable to identify patients that are most likely to benefit from antibiotics when applying conventional (*i.e.*, univariable or one-variable-at-a-time) subgroup analysis. We investigated whether multivariable prediction of patient-level prognosis and antibiotic treatment effect may lead to more tailored treatment assignment in adults presenting to primary care with ARS. An IPD-MA of nine double-blind placebo-controlled trials of antibiotic treatment (n=2,539) was conducted, with the probability of being cured at 8-15 days as the primary outcome. A logistic mixed effects model was developed to predict the probability of being cured based on demographic characteristics, signs and symptoms, and antibiotic treatment assignment. Predictive performance was quantified based on internal-external cross-validation in terms of calibration and discrimination performance, overall model fit, and the accuracy of individual predictions. Results indicates that the prognosis with respect to risk of cure could not be reliably predicted (c-statistic 0.58 and Brier score 0.24). Similarly, patient-level treatment effect predictions did not reliably distinguish between those that did and did not benefit from antibiotics (c-for-benefit 0.50). In conclusion, multivariable prediction based on patient demographics and common signs and symptoms did not reliably predict the patient-level probability of cure and antibiotic effect in this IPD-MA. Therefore, these characteristics cannot be expected to reliably distinguish those that do and do not benefit from antibiotics in adults presenting to primary care with ARS.

6.1 Introduction

Acute rhinosinusitis (ARS) is one of the conditions with highest antibiotic over-prescription rates in adults [166, 167]. With antimicrobial resistance posing a serious threat to global public health [168], continuous efforts are needed to reduce inappropriate antibiotic prescription in primary care [169]. One of the reasons for the persistent habit of general practitioners (GPs) to prescribe antibiotics might be attributed to their clinical impression that there is a subgroup of patients with clinically diagnosed ARS that actually do benefit from antibiotics [170]. There is also some evidence to substantiate this impression; antibiotics seem to have larger effects in those with radiologically confirmed ARS, in particular those with a fluid level or total opacification in any sinus on computed tomography [171]. Previous attempts to identify these subgroups based on common signs and symptoms were not successful, including an individual patient data meta-analysis (IPD-MA) of randomized controlled trials (RCTs) comparing antibiotics with placebo in adults with clinically diagnosed ARS [172]. This preceding IPD-MA applied conventional (univariable) subgroup analysis in which potential effect modification of single signs and symptoms was assessed one at the time. This approach does not focus on the absolute risk scale that is of most interest for clinical decision making (instead focusing on relative effects), likely under-represents underlying clinical heterogeneity (individuals may vary in more than one relevant aspect) [89, 173], and is known to be statistically inefficient [174]. Multivariable risk prediction modelling allowing for simultaneous analysis of multiple baseline variables that may influence treatment effect has the potential to overcome these problems [173, 175, 90, 133, 146]. Such a model provides patient-level outcome risk predictions for both treatment assignments and hence also predicts the patient-level absolute benefit of antibiotic treatment of interest. Due to the required sample size, IPD from multiple studies provide a good source for model development [176, 177]. Subsequently if accurate predictions can be made, they can inform treatment decisions in clinical practice, informing on the probability of fast spontaneous resolution of symptoms and the anticipated benefit of antibiotic treatment at the patient-level. With this aim, we applied multivariable prediction modelling methods to IPD of multiple RCTs comparing antibiotics with placebo in adults with clinically diagnosed ARS.

6.2 Methods

The protocol of this IPD-MA has been registered in PROSPERO (registration number CRD 42020220108) and published [178]. A detailed description of the rationale and methodology can be found in the protocol publication [178]. We followed recommendations provided in the Predictive Approaches to Treatment effect Heterogeneity (PATH) statement [90], guidance on the individualized treatment effect prediction [146], and guidance on the use of IPD-MA of diagnostic and prognostic modelling studies [177], and reported according to the TRIPOD [179, 180] and PRISMA-IPD statement [181].

6.2.1 Study identification and selection

We conducted a systematic search to identify eligible studies. First, the reference list of the 2018 Cochrane review on antibiotics for ARS in adults [171] was reviewed for any relevant studies published since the 2008 IPD-MA [172]. Next, we updated the systematic electronic searches of the Cochrane review (online supplementary Table S1) from January 18, 2018 (date of last search) to September 1, 2020 to increase the yield of potentially relevant trials. No language restrictions were applied.

Titles and abstracts of the unique records retrieved from these electronic databases were screened and the full text of all potentially eligible articles was reviewed against the following predefined criteria: (i) RCT comparing antibiotics with placebo, and (ii) enrolled adults ($\geq$ 16 years) presenting to primary care with uncomplicated ARS based on clinical signs and symptoms. Studies involving children ($<$ 16 years), referred patients, hospitalized patients as well as those involving highly specialized populations (*e.g.*, those with immunodeficiency, odontogenic sinusitis, or malignancy) were excluded. In addition, reference lists of all eligible studies as well as those from relevant systematic reviews were screened for any further potential studies and contributing review authors were asked if they knew any additional (published or unpublished) studies. Study authors of eligible trials were contacted and invited to provide the de-identified, complete dataset of their original trial.

6.2.2 Quality assessment of included studies

Methodological quality of the included studies was assessed using the Cochrane Risk of Bias 2 tool [182]. If information regarding study quality was unclear or

undisclosed, individual trial authors were contacted to provide further clarification.

6.2.3 Outcome assessment

All retrieved IPD were assembled in a single dataset. The predefined outcome of interest was cure at 8-15 days (yes vs no) [178], which was available in all studies.

6.2.4 Candidate predictors

Candidate predictors were selected based on clinical reasoning, knowledge from existing literature and availability in the IPD set. Next to (i) treatment assignment (oral antibiotics vs placebo) which was available in all trials, the following pre-specified candidate predictors of treatment effect were available in at least 50% of studies: (ii) sex, (iii) age (in years), (iv) preceding upper respiratory tract infection (URTI), (v) symptom duration prior to enrolment (in days), (vi) pain on bending, (vii) teeth pain, (viii) unilateral facial pain, (ix) self-reported purulent nasal discharge (PNDsr), (x) symptom severity, (xi) presence of fever ($>$ 37.5 C; yes vs no), (xii) purulent nasal discharge upon examination (PNDex), and (xiii) purulent pharyngeal discharge upon examination (PPDex). For symptom severity, we used the standardized 0-100 severity as used in the 2008 IPD-MA [172] which was based on a (scaled) logistic transformation of the severity measures applied in the individual trials. The following pre-specified candidate predictors [178] could not be included in our analysis due to not being measured in $>$ 50% of trials: previous ARS, anosmia, cacosmia, double sickening, overall clinical impression, C-reactive protein (CRP), and erythrocyte sedimentation rate (ESR) values.

6.2.5 Sample size considerations

We calculated the maximum number of candidate predictors based on an anticipated number of 2500 patients in the IPD set, with an average outcome prevalence of 60% cure, and a desired 0.05 accuracy in terms of mean absolute prediction error [103]. Since the available guidance does not yet extend to clustered IPD, we conservatively estimated our effective sample size to be 1250 which allows for evaluation of 25 parameters in the model based on a presumed Cox-Snell R^2 of 0.175, which is also expected to keep shrinkage below 10% and the expected Cox-Snell R^2 within 5%.

6.2.6 Statistical analysis

Handling of missing data

Missing data were imputed using a fully Bayesian joint modelling approach [18]. A total of 50 imputations were derived as compatible with a generalized linear mixed effects analysis model with a logistic link function, random intercepts per study, main effects for treatment and each of the candidate predictors, and treatment-predictor interaction terms [17]. All effect were modelled to be linear on the linear predictor scale since spline-based exploratory analysis based on the complete cases did not indicate clear non-linear predictor-outcome relations.

Descriptive statistics

First, predictors and outcome distributions were summarized in each study. Next, a multinomial membership model was used to evaluate multivariable between-study heterogeneity in predictor and outcome distributions [183]. Such a membership model predicts study membership based on the candidate predictors and outcome and hence illustrates the degree to which multivariable differences between studies allow a model to predict to which study an individual belongs. Details are provided in the online supplementary material 1.

Main analysis: prediction model development

In the primary analysis, all available candidate predictors and treatment assignment were included as main effects in a logistic mixed effects regression model with random intercepts per study [178]. Symptom duration was heavily skewed to the right and therefore log-transformed. Due to between-study variability in outcome assessment, study level variables 'number of days between baseline and outcome measurement' and 'type of outcome measurement' were added to the model. In absence of strong evidence for pre-specification of certain treatment-predictor interactions, we included treatment-predictor interactions for all available predictors and performed a pooled likelihood ratio test (based on the D_3-statistic [17]) for their combined significance as a conservative approach [178].

Secondary and exploratory analyses

As opposed to the study of individual treatment interactions, baseline risk-modelling [90] was pre-specified as a secondary analysis [178]. This approach

entails an evaluation of possible treatment effect heterogeneity as a function of baseline risk-model, and has been recommended in settings where i) an overall treatment effect is well established, ii) several large RCTs are available for analysis, and iii) when substantial identifiable heterogeneity of outcome risk in the trial population(s) is anticipated [90]. In addition, in order to evaluate the possible benefit of model simplification in terms of generalizability, model reduction was evaluated using a relaxed-lasso procedure in exploratory analysis [184, 185]. The relaxed-lasso was performed on stacked imputed data [186], with fixed and unpenalized study intercepts, an unpenalized main treatment effect, and penalized main effects for all candidate predictors, and penalized interactions between all candidate predictors and treatment. Tuning parameters lambda (degree of penalization selection) and gamma (degree of post-selection relaxation) selected according to the 1 standard error rule based on 10-fold cross-validation.

Evaluation of prediction model performance

Prediction model performance with respect to the prediction of outcome risk and absolute antibiotic treatment effect was evaluated by means of calibration performance (extent of agreement between predicted risk and observed events), discrimination performance (with the aim to quantify whether predicted risk correctly rank-orders actual risk), Nagelkerke R^2 (as a measure of overall model fit) and Brier score (as a measure of prediction accuracy). Performance was assessed using internal-external cross-validation (IECV) [33]. Standard errors for each of the measures were derived based on 500 bootstrap samples. Meta-analysis was used to summarise the main IECV results using restricted maximum likelihood-based estimates of between study variability, inverse variance weighting, and Hartung and Knapp adjustment [187]. Prediction model performance with respect to predicted absolute antibiotic treatment effect (*i.e.*, on the risk difference level) was evaluated in terms of discriminative performance using the c-for-benefit [132] and in terms of calibration in the form of predicted versus observed treatment effect in quartiles of predicted treatment effect.

6.3 Results

6.3.1 Study inclusion and study characteristics

The 2008 IPD-MA [172] included data from 9 trials [188, 189, 190, 191, 192, 193, 194, 195, 196]. An additional eligible study [197] was identified from reviewing the reference list of the 2018 Cochrane review [171]. This study with 166 par-

ticipants (online supplementary Table S2) was excluded since authors were not able to provide IPD. No further eligible studies were found after screening the 303 unique records retrieved from the electronic database searches or through additional routes (Figure 6.1). This left 9 trials with 2,539 participants aged ($\geq$ 16 years) for inclusion [188, 189, 190, 191, 192, 193, 194, 195, 196]. Details on the design characteristics of the included studies are shown in online supplementary Table S3. All studies were double-blind, placebo controlled randomised trials and conducted in high-income countries in Europe and in the US. One trial used a 2x2 factorial design [195], and data were split into two subtrials: antibiotics vs. placebo without concomitant nasal steroids in both groups (Williamson1) or antibiotics vs. placebo with concomitant nasal steroids in both groups (Williamson2). Participants from the intervention groups received beta-lactam antibiotics (mainly amoxicillin, but also amoxicillin clavulanate or phenoxymethylpenicillin), macrolides (azithromycin), or tetracyclines (doxycycline). Sample size of the included trials ranged from 135 to 503.

6.3.2 Quality assessment of included studies

The quality assessment of included studies is summarised in online supplementary Figure S1. The risk of bias could not be assessed for the unpublished Schering-Plough trial [196]. Overall risk of bias was judged low for the other included studies.

6.3.3 Missing data

The percentage of missing data varied greatly across studies and variables (online supplementary Table S4). Including both sporadic (*i.e.* partly, but not completely missing in a certain study) and systematically missing data (*i.e.* completely missing in a certain study), the percentage of missingness was below 10% for all variables except for preceding URTI (66%, unavailable in 5/10 studies) pain on bending (62%, unavailable in 5/10 trials), pain in teeth (56%, unavailable in 4/10 trials), unilateral facial pain (41%, unavailable in 2/10 trials), and PPDex (52%, unavailable in 5/10 trials).

6.3.4 Descriptive statistics

Descriptive statistics for each of the trials after imputation of missing data are shown in Table 6.1 and visually presented in online supplementary Figure S2. Studies differed with respect to both outcome occurrence (range 35-77%) and

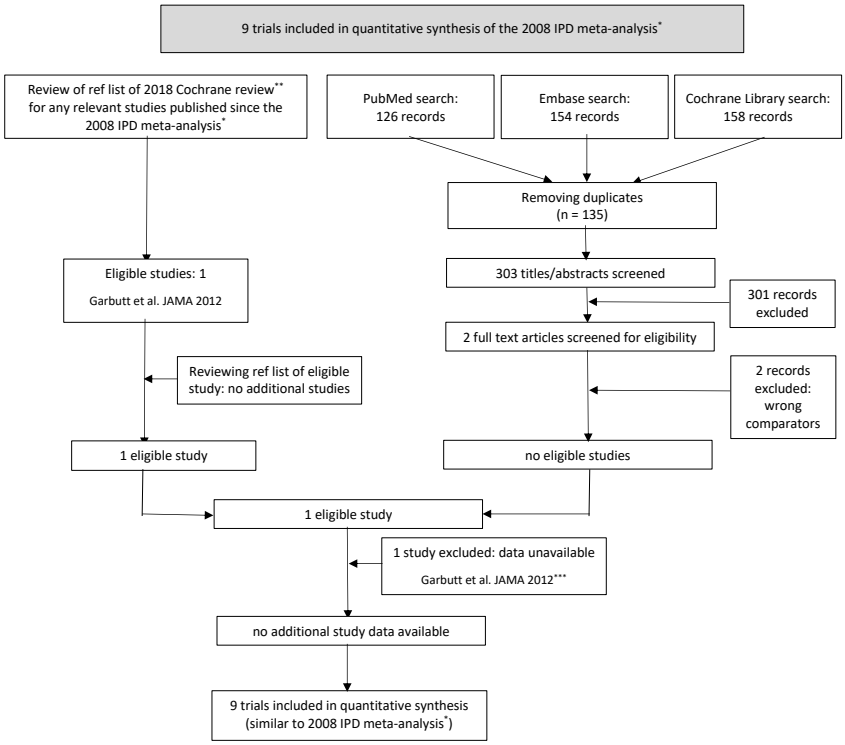


Figure 6.1: Inclusion flowchart. * refers to Young et al. [172], ** refers to Lemiengre et al. [171], and *** refers to Garbutt et al. [197].

the prevalence of predictors of interest. Most notably, symptom duration prior to enrolment, and the prevalence of pain on bending, PNDsr, and PPDex varied substantially across studies.

Trial	Bucher	De Sutter	Kaiser	Meltzer	Merenstein
Number of participants	251	368	266	490	135
Antibiotic assignment (%)	49	49	51	50	50
Female (%)	54	57	52	66	69
Age in years, mean (SD)	36.6 (13.1)	38.4 (14.2)	32.8 (11.6)	35.7 (12.2)	33.9 (9.8)
Preceding URTI (%)	56	79	35	78	52
Symptom duration in days median (IQR)	4 (3,7)	6 (3,10)	4 (3,7)	13 (10,19)	9 (7-14)
Pain on bending (%)	82	69	27	86	81
Pain in teeth (%)	24	44	14	85	28
Unilateral facial pain (%)	54	55	14	98	39
PNDsr (%)	98	88	51	67	84
Symptom severity, median (IQR)	44.2 (19.8, 67.8)	50.6 (24.4, 62.7)	46.0 (15.8, 74.1)	41.8 (23.4, 79.6)	43.7 (27.0, 78.6)
Fever (%)	14	7	11	1	1
PNDex (%)	66	55	33	33	32
PPDex (%)	21	45	18	45	55
Cure (%)	69	35	56	69	49

(Continued)	**Schering-Plough**	**Stalman**	**Varonen**	**Williamson1**	**Williamson2**
Number of participants	458	191	150	116	114
Antibiotic assignment (%)	48	51	59	51	45
Female (%)	70	64	70	72	76
Age in years, mean (SD)	38.7 (14.6)	37.0 (11.4)	39.7 (13.4)	43.3 (15.5)	43.8 (12.7)
Preceding URTI (%)	76	92	56	71	67
Symptom duration in days, median (IQR)	13 (10,19)	7 (6,14)	5 (3,9)	5 (3,10)	5 (3,7)
Pain on bending (%)	85	90	78	82	85
Pain in teeth (%)	81	50	19	37	56
Unilateral facial pain (%)	98	42	23	52	51
PNDsr (%)	53	83	48	74	67
Symptom severity, median (IQR)	48.5 (28.4, 77.5)	48.1 (23.1, 74.7)	55.7 (25.7, 76.0)	56.2 (27.7, 82.1)	48.5 (27.7, 73.1)
Fever (%)	2	4	18	3	4
PNDex (%)	31	12	32	39	31
PPDex (%)	34	9	72	22	25
Cure (%)	66	65	77	64	65

Table 6.1: Descriptive statistics per trial after multiple imputation
IQR: interquartile range; PNDex: purulent nasal discharge upon examination; PNDsr: purulent nasal discharge self-reported; PPDex: purulent pharyngeal discharge upon examination; SD: standard deviation; URTI: upper respiratory tract infection. Grey figures relate to systematically missing data.

Online supplementary Table S5 further illustrates the between-study heterogeneity. The membership model had high discriminative ability for all studies, indicating substantial differences in predictor and outcome distributions across studies. Based on a common intercept and common predictor-outcome associations, the observed outcome incidence deviated somewhat from the expectation for four trials (Merenstein et al. [194], Kaiser et al. [189], de Sutter et al. [190], and Varonen et al. [192]), indicating that the observed incidence of cure could not be completely explained by the modeled effects of case-mix differences (online supplementary Figure S3).

6.3.5 Main analysis results

Estimates for the pre-specified main effects model are shown in Table 6.2. Significant patient-level associations with the risk of cure were found for antibiotic treatment (OR 1.34 [1.13 to 1.59]), age (OR 0.91 per 10 years [0.85 to 0.97]), log symptom duration prior to enrolment (OR 0.76 [0.65 to 0.89]), symptom severity (OR 0.87 [0.82 to 0.91]). A significant study-level association with the risk of cure was found for outcome assessment based on clinical examination or a combination of methods vs. symptom diary (OR 0.40 [0.19 to 0.84]). Despite these main effect estimates, there was still considerable unexplained between-study variability in the outcome as shown in the random intercepts estimates (online supplementary Figure S4). The estimated standard deviation of the random intercept distribution was 0.33, with the largest deviations for study data from Merenstein (-0.48), de Sutter (-0.43), Kaiser (0.44), and Varonen (0.48). All treatment-predictor interactions were dropped from the model based on the planned pooled likelihood ratio test of their combined contribution (D3 statistic 0.54, df_1 12, df_2 7497, p = 0.89).

	est $\hat{\beta}$	**se**	**95% CI**	**OR (95% CI)**
Intercept	0.83	0.88	(-0.89, 2.56)	2.30 (0.41, 12.98)
Antibiotics (yes)	0.29	0.09	(0.12, 0.46)	1.34 (1.13, 1.59)
Sex, female	-0.09	0.09	(-0.27, 0.09)	0.92 (0.76, 1.10)
Age, per 10 years	-0.10	0.03	(-0.16, -0.03)	0.91 (0.85, 0.97)
Preceding URTI	0.23	0.19	(-0.15, 0.61)	1.26 (0.86, 1.84)
Symptom duration in log(days))	-0.27	0.08	(-0.42, -0.12)	0.76 (0.65, 0.89)
Pain on bending	0.12	0.18	(-0.23, 0.47)	1.13 (0.80, 1.60)
Pain in teeth	-0.12	0.15	(-0.42, 0.18)	0.89 (0.66, 1.20)
Unilateral facial pain	0.14	0.13	(-0.11, 0.40)	1.15 (0.89, 1.49)
PNDsr	0.17	0.11	(-0.05, 0.39)	1.19 (0.95, 1.48)
Symptom severity*	-0.14	0.03	(-0.20, -0.09)	0.87 (0.82, 0.91)
Fever	-0.25	0.19	(-0.63, 0.13)	0.78 (0.53, 1.14)
PNDex	0.07	0.10	(-0.12, 0.26)	1.07 (0.89, 1.29)
PPDex	-0.19	0.15	(-0.48, 0.10)	0.82 (0.62, 1.10)
Time to outcome measurement (days)	0.03	0.08	(-0.12, 0.18)	1.03 (0.89, 1.20)
Outcome type: other, (ref. diary)	-0.92	0.38	(-1.66, -0.18)	0.40 (0.19, 0.84)
Outcome type: telephone, (ref. diary)	-0.13	0.37	(-0.85, 0.59)	0.88 (0.43, 1.81)

Table 6.2: Main effect estimates for the random intercept model. Coefficients (log(OR)), standard errors, odds ratios (OR) and 95% confidence intervals (CI) as pooled across imputations. The mean standard deviation of the random intercepts was 0.33.
*) per point on the inverse logit transformation of (severity score / 100).
Abbreviations: PNDex purulent nasal discharge upon examination; PNDsr purulent nasal discharge self-reported; PPDex purulent pharyngeal discharge upon examination; URTI upper respiratory tract infection.

IECV performance estimates indicated poor prediction performance and overall model fit of the model main effects model (Table 6.3 and online supplementary Figure S5). The pooled IECV c-statistic estimate (0.58) did indicate some discriminative ability with a prediction interval (PI) of 0.56-0.62. However, while R^2 and Brier scores were heterogeneous across studies, their pooled estimates clearly indicate poor performance with R^2 -0.08 (PI -0.48, 0.32) and Brier score 0.24 (PI 0.15,0.34). Both measures indicate that the main effects model did not provide accurate absolute risk predictions for the hold-out studies. This lack of generalizability between studies was further illustrated by the large prediction intervals for the estimated calibration intercepts [-1.06 and 1.11] and calibration slopes [0.18 and 1.38]. While these intervals include the favourable values of 0 and 1, they also include a large range of unfavourable calibration estimates.

As a sensitivity analysis, all analyses were re-run after omitting data from the Schering-Plough study [196], as the risk of bias could not be assessed for this trial. This, however, did not substantially change model performance (online supplementary Table S6). In summary, the absolute risk of cure could not be reliably predicted based on the available predictors, and can hence not be used to differentiate between low-risk and high-risk individuals to inform treatment decisions.

	C-statistic	**R^2**	**Brier**	**Intercept**	**Slope**
Bucher	0.59	0.02	0.21	0.08	0.79
De Sutter	0.55	-0.49	0.30	-0.82	0.38
Kaiser	0.55	-0.60	0.33	0.61	0.38
Meltzer	0.62	0.03	0.21	0.00	1.57
Merenstein	0.57	-0.26	0.29	-0.51	0.57
Schering-Plough	0.58	0.02	0.22	0.15	0.88
Stalman	0.59	0.04	0.22	-0.04	0.92
Varonen	0.62	-0.27	0.21	0.94	1.09
Williamson1	0.60	0.01	0.23	0.09	0.67
Williamson2	0.62	0.04	0.22	-0.10	1.04

Table 6.3: IECV results for risk (of cure) prediction based on the main effects and random intercept model.
IECV: internal-external cross-validation.

6.3.6 Secondary and exploratory analyses results

Given the lack of reliable risk predictions based on the main risk model, further modelling using these predictions as inputs was not deemed relevant. Therefore, baseline risk-modelling, which essentially evaluates outcome risk modification by treatment, was not performed. As anticipated based on previous findings, the exploratory relaxed-lasso procedure led to substantial model reduction: only a main effect for symptom severity and unpenalized parameters (study intercepts and treatment assignment) were left in the model.

Contrary to the large between-study heterogeneity in terms of model performance as observed in the main analysis, evaluation of the marginal relative treatment effect (OR 1.32; 95% CI 1.11-1.56) did not reveal any between-study heterogeneity (not shown), confirming earlier results[172].

6.3.7 Evaluations of absolute treatment effect prediction

To supplement outcome risk evaluations, individual predictions of absolute treatment effect were evaluated (online supplementary Figure S6). The IECV estimate for discriminative performance (c-for-benefit) was 0.50 for the main effects model, indicating absence of discriminative ability. Therefore, further examination of calibration performance was not deemed relevant.

6.4 Discussion

This large IPD-MA of high-quality antibiotic therapy trials in adults presenting to primary care with clinically diagnosed uncomplicated ARS evaluated patient-level variability in prognosis and antibiotic treatment effect. Such variability could not be reliably predicted based on demographics and clinical signs and symptoms, illustrating that these characteristics do not contribute to the identification of patients that are most likely to benefit from antibiotics.

A strong aspect of this study was the large sample size derived from multiple high-quality trials. This allowed for careful handling of missing data and consistent multivariable prediction modelling of antibiotic treatment effect across studies [177]. The lack of predictable between-subject heterogeneity of antibiotic benefit was robust, since our conservative primary analysis' findings were supported by those derived from exploratory relaxed-lasso modelling.

Several limitations deserve further attention. First, we observed a high degree

of heterogeneity across studies, in particular with respect to the outcome definition, outcome assessment, and studied populations. In terms of outcome definition and assessment, this was alleviated by adjustment for study level information on time to outcome assessment and type of outcome assessment. With respect to heterogeneity in study populations, internal-external cross-validation revealed that a common model did not describe the data well. Second, we did not have sufficient information to include time-to-cure instead of the available dichotomous outcome data, which would likely be a more sensitive outcome. Also, severely unwell individuals with prolonged illness duration may be underrepresented in the included trials, and the modeled relationships between predictors and outcome may not generalize to the wider population presenting in primary care. Third, there was a substantial amount of systemically missing data. Although carefully handled using multiple imputation, this still represents loss of information which likely has influenced our results (e.g. possibly weakening predictor-outcome associations). Finally, potential important signs (severe pain, double sickening), and laboratory findings (CRP, ESR) were not available in a sufficient number of trials. It is, however, uncertain whether the availability of these variables would have impacted our findings. For example, CRP was found to be of value in a recent diagnostic IPD-MA for ruling out, but not for ruling in target conditions associated with antibiotic benefit in adults suspected of ARS [198]. A recent review of diagnostic accuracy studies of CRP, ESR, white blood cell counts, procalcitonin, and nasal nitric oxide for detecting acute bacterial rhinosinusitis (ABRS) found that especially elevated CRP and ESR are associated with higher probability of ABRS. However, CRP and ESR were still found insufficiently accurate for predicting ABRS [199]. Further research in this field should focus on the added value of novel point-of-care tests or novel devices such as those aimed at gaining specimens from draining sinuses [200] over readily available signs and symptoms such as age, symptom duration and severity. Early-stage investigations of biomarker combination tests as well as host gene expression diagnostics suggest that these point-of-care tests have the potential to discriminate between viral and bacterial aetiology of RTI, but high-quality prospective clinical validation studies in primary care are needed to confirm their potential [201, 202, 203].

In conclusion, this IPD-MA did not find evidence to support predictable heterogeneous antibiotic treatment effect based on demographics and signs and symptoms in adults presenting to primary care with clinically diagnosed ARS. While future research may reveal markers that aid the identification of adults with clinically diagnosed ARS most likely to benefit from antibiotics, current

evidence does not support individualized treatment selection in adults with uncomplicated ARS.

Acknowledgements

This study was supported by The Netherlands Organisation for Health Research and Development (grant 91618026). The funder did not participate in the design of the study and will have no role in the study conduct, data analysis, interpretation, and publication of the data.

Supplementary Material

The supplementary material for this chapter was too extensive to fully include in print and is available online (`https://github.com/jeroenhoogland/ARS_Abx_IPDMA_suppl`)

Chapter 7

General discussion

"Prediction is very difficult, especially about the future."

Danish proverb, often attributed to Niels Bohr [204]

The chapters of my dissertation have all focused on prediction modeling and are part of a large body of literature on this topic. Nonetheless, the cited epigraph goes back to at least the 1930's [204], and remains telling today. This general discussion aims to shortly touch upon two overarching themes that connect the chapters of this dissertation, while shedding some light on remaining difficulties and possible future directions for medical prediction modeling.

Dealing with the unknown; what are we missing?

> "Your assumptions are your windows on the world. Scrub them off every once in a while, or the light won't come in."
>
> *Alan Alda*

In a sense, all chapters revolved around incompletely observed data or incompletely observable processes. Particular problems touched upon include partially unobserved covariate data (**Chapter 2**), partially observed or censored outcome data (**Chapter 3**), and the unobserved nature of all but one potential outcome per individual in intervention research (**Chapters 4, 5 & 6**). The key question is whether we can find a way for the observed data to form a plausible representation of the complete picture. The only way forward is to formulate plausible assumptions about the relations between the observed and the unobserved data or processes.

The degree to which the data provide information on the unknowns differs greatly. In fact, many assumptions cannot be directly assessed based on the data. For instance, in presence of missing data, the validity of the missing at random assumption (with missing not at random being the alternative) cannot be assessed based on the data (**Chapter2**). The same holds for the assumed conditional independence between the event time distribution and censoring distribution in survival analysis (**Chapter 3**). Likewise, the identifiability assumptions supporting causal inference cannot be readily verified based on just data (**Chapters 4, 5 & 6**). In some instances, the study design provides plausibility to the assumptions, such as for randomized studies and the exchangeability assumption. Other times, conceptual knowledge of the underlying processes being modeled is a major factor. These instances may require careful thought regardless of the amount of data.

Other assumptions are more easily checked if a sufficient amount of data is available. For instance, data are a great help in presence of weak prior knowledge with respect to model specification in terms of functional form or model sparsity. The challenge here is that data are often available in limited amount, whether for pragmatic or more insurmountable reasons. This particular problem was approached using regularization in **Chapter 3**, but a wide range of methods is available (*e.g.*, as introduced in [20]). Even here though, careful thought on

the method of choice is key and should match the specific research problem at hand. For instance, one can try to choose between ridge and lasso penalization purely based on the data, but it might be more fruitful to decide on an analysis strategy based on the expected degree of sparsity for the particular problem [61]. Similarly, one can try to choose between a tree-based algorithm and regression based on a resampling algorithm, but careful thought on the expected degree of non-linearity and interaction may be more revealing in settings with limited data. The problem is that a particular sample of limited size often provides inconclusive evidence with respect to the best choice of modeling strategy. Breiman referred to this problem as model instability: a small perturbation in the data may lead to a rather different model with very similar performance [205]. In practice, the combination of limited theoretical knowledge and limited sample size is the rule rather than the exception in the medical domain and leads to a high degree of uncertainty about the optimal modelling strategy that best answers the research question. The ensuing models should only form a starting point for an iterative scientific process that enables learning. However, this may not always be sufficiently recognized. In addition, the process being modeled may change over time, which necessitates monitoring and updating of a model over time [206].

Lastly, a common argument in prediction research is that knowledge of the inner workings of the prediction model are not very important as long as it predicts well. That is, the prediction procedure might just as well be a black box. This may hold in fields where generalizability or transportability of models is not of major interest, and where the model of interest can easily be checked in the situation in which it will be used. However, in the medical domain, models will always be applied in settings that differ from the development setting. A problem with black-box models is that it is more difficult to gauge the expected performance in a novel but slightly different setting [207]. In contrast, if a model's structure is based on understanding, it is more likely to generalize well in the first place, and its structure will provide clues on the to be expected performance in a new domain with specific characteristics. Also, while the black-box problem is more evident for machine learning methods, even simple models may reflect spurious associations and likewise fail to generalize in unanticipated ways. A better understanding of the causal roles of all variables in a model would allow for improved reasoning with respect to generalizability, as is possible for many models in physics. Vice versa, better causal model may help to improve understanding of the underlying processes. However, as of yet, the underlying biological processes in the medical domain are often too complex for

such details understanding, and the subsequent reliance on association learning is not without dangers. This is not to be taken as a problem per se, as long as it is recognized as a starting point and not as an endpoint.

In summary, much progress has been made in terms of the statistical modeling of data for the purpose of prediction, but data often do not tell the whole story. Extrapolation from data often relies heavily on assumptions, and the validity of these assumptions may be challenging to assess. While external data provide a powerful means to independently assess many modeling assumptions, some assumptions remain more elusive and require more fundamental understanding of the process being modelled.

Association learning versus causal prediction

Classically, prediction models aim to model an outcome of interest as based on its association with a certain set of measured variables. As already mentioned, it may be useful to know more about the function of those measured variables, but when the purpose is just to predict in a well-known setting, it may be sufficient to accurately reflect the associations in the data. The first two chapters focused on this type of prediction modeling. In contrast, when the model is used to inform some intervention on one of more variables, interest is in the variability in the outcome that is *caused* by these variables. This was the main topic of the remaining chapters.

In the end, many research questions of interest are causal. Indeed, questions about causation are as old as science itself and not only relate to interventions but also to understanding. In statistics, early work was performed in the twenties by Neyman and later in the eighties by Rubin [23]. Recent decades have shown a revived interest in causal inference in the fields of statistics, econometrics, and machine learning (*e.g.* [208, 21, 22]).

The fundamental difference between pure prediction (*e.g.*, weather forecast, natural course of disease) and causal prediction (*e.g.*, about the effect of a certain intervention) is increasingly acknowledged. In parallel, both fields have developed rapidly. While this dissertation has primarily focused on causal prediction from randomized trial data, a large body of literature exists on causal inference in observational data (a recent overview is provided elsewhere [24]). The key challenge in observational data is to achieve exchangeability with respect to the causal variable of interest, which requires some form of adjustment for

confounders and selection processes. Also, causal inference work is being done in the context of high-dimensional data (*e.g.* [209]).

In the context of medical statistics and medical practice, there are many examples of the limitations of associative prediction modeling that are not always recognized. For instance, it is a lapse of judgement to check the influence associated with a change in an important covariate such as smoking in an associative model predicting cardiovascular disease, and subsequently conclude that this reflects the effect of intervening on smoking habits. The 'table 2 fallacy' (*i.e.,* the misconception that all effects in a table of multivariate-adjusted associations with the outcome can be interpreted in a simular manner) is another example of the same underlying confusion [210]. In similar fashion, as illustrated in **Chapter 3**, observing that treatment effect varies with some marker does not imply that the marker causes this variation. It has been argued that explicitly modeling the causal effects of treatment policies may render prediction models more robust to changing policies over time [130]. This includes the increasingly recognized case of prognostic models being the 'victims of their own success' [211]. Also, as alluded to above, work has been done on the conditions under which model transportability could be expected [212].

While the promises of structural causal models [21, 22] are enticing, their successful implementation again requires a substantial level of content knowledge to feed into the modeling process with the right assumptions. In the medical domain, the knowledge of the processes being modeled is often limited, and data are frequently used with the aim to increase understanding from the bottom up. That is, modeling is not used to test scientific theories, but to generate new ones. Likewise, prediction modeling is often approached as a bottom-up process in the medical domain. Here, the aim is just to distill accurate predictions from the data, and the test is whether the model performs well in external data. While this may be sufficient for some purposes, it ultimately does not increase understanding if not followed up by additional investigation of the underlying processes. If anything, the structure required for causal inference necessitates attention to the role that different variables play, and hence for explicit thoughts about confounders, selection processes, possible mediators, and potential outcome definitions. As content knowledge grows, classically associative prediction models may be increasingly endowed with known structures and hence improve accuracy, reduce spurious association, increase transportability, and ultimately increase understanding. The paradox is that knowledge is required for more efficient learning.

Concluding remarks

The chapters in this dissertation presented recent developments and guidance in the field of clinical prediction modeling, with a specific focus on missing data, regularization, and causal inference. While these fields have largely developed independently, many interesting research questions can be found on their intersection, and the synergy between prediction methods and causal inference techniques seems especially promising. The humble ambition with respect to the impact of this work is to increases awareness of the issues surrounding these topics and provides some guidance when encountering these topics during prediction model development, implementation, and evaluation in practice.

Bibliography

[1] Friedman JH. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery. 1997;1:55-77.

[2] Shmueli G. To Explain or to Predict? Statistical Science. 2010;25(3). Available from: https://projecteuclid.org/journals/statistical-science/volume-25/issue-3/To-Explain-or-to-Predict/10.1214/10-STS330.full.

[3] Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. Circulation. 1998 May;97(18):1837-47. Available from: https://www.ahajournals.org/doi/10.1161/01.CIR.97.18.1837.

[4] D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, et al. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation. 2008 Feb;117(6):743-53. Available from: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.107.699579.

[5] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017 May;357:j2099. Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.j2099.

[6] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. JNCI Journal of the National Cancer Institute. 1989 Dec;81(24):1879-86. Available from: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/81.24.1879.

[7] Pfeiffer RM, Gail MH. Absolute Risk: Methods and Applications in Clinical Management and Public Health. 1st ed. Boca Raton : Taylor & Francis, a CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa plc, 2017. |: Chapman and Hall/CRC; 2017. Available from: https://www.taylorfrancis.com/books/9781466561687.

[8] Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. Springer series in statistics. Cham Heidelberg New York: Springer; 2015. OCLC: 922304565.

[9] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Statistics for Biology and Health. Cham: Springer International Publishing; 2019. Available from: http://link.springer.com/10.1007/978-3-030-16399-0.

[10] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012 May;98(9):683-90. Available from: http://heart.bmj.com/lookup/doi/10.1136/heartjnl-2011-301246.

[11] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012 May;98(9):691-8. Available from: http://heart.bmj.com/lookup/doi/10.1136/heartjnl-2011-301247.

[12] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Statistics in Medicine. 2019 Mar;38(7):1262-75. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.7993.

[13] Riley RD, Snell KI, Ensor J, Burke DL, Harrell Jr FE, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019 Mar;38(7):1276-96. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.7992.

[14] Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. Statistics in Medicine. 2021 Jan;40(1):133-46. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8766.

[15] Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2019.

[16] White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377-99. Available from: http://doi.wiley.com/10.1002/sim.4067.

[17] Buuren Sv. Flexible imputation of missing data. 2nd ed. Chapman and Hall/CRC interdisciplinary statistics series. Boca Raton: CRC Press, Taylor & Francis Group; 2018.

[18] Erler NS, Rizopoulos D, Lesaffre EMEH. JointAI: Joint Analysis and Imputation of Incomplete Data in R. arXiv:190710867 [stat]. 2020 Sep. ArXiv: 1907.10867. Available from: http://arxiv.org/abs/1907.10867.

[19] Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics. 2018 Sep;21(2):236-52. Available from: https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy040/5092384.

[20] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer series in statistics. New York: Springer; 2017. OCLC: 995842694.

[21] Pearl J, Glymour M, Jewell NP. Causal inference in statistics: a primer. Chichester, West Sussex: Wiley; 2016.

[22] Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.

[23] Holland P. Statistics and Causal Inference. Journal of the American Statistical Association. 1986;81(396):945-60.

[24] Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagnostic and Prognostic Research. 2021 Dec;5(1):3. Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-021-00092-9.

[25] Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle Heart Failure Model: Prediction of Survival in Heart Failure. Circulation. 2006 Mar;113(11):1424-33. Available from: https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.105.584102.

[26] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Medicine. 2013 Feb;10(2):e1001381. Available from: http://dx.plos.org/10.1371/journal.pmed.1001381.

[27] Buuren Sv. Flexible imputation of missing data. Boca Raton, FL: CRC Press; 2012. OCLC: 690084672.

[28] Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: an overview and case study. Emerging Themes in Epidemiology. 2017 Dec;14(1). Available from: http://ete-online.biomedcentral.com/articles/10.1186/s12982-017-0062-6.

[29] Buuren Sv, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1-67.

[30] Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. BMC Medical Research Methodology. 2014 Dec;14(1):116. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-116.

[31] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009 May;338:b605. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.b605.

[32] Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine. 2000 Feb;19(4):453-73. Available from: http://doi.wiley.com/10.1002/%28SICI%291097-0258%2820000229%2919%3A4%3C453%3A%3AAID-SIM350%3E3.0.CO%3B2-5.

[33] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology. 2016 Jan;69:245-7. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0895435615001754.

[34] Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015 Mar;68(3):279-89. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435614002753.

[35] Schomaker M, Heumann C. Bootstrap inference when using multiple imputation: Bootstrap Inference When Using Multiple Imputation. Statistics in Medicine. 2018 Jun;37(14):2252-66. Available from: http://doi.wiley.com/10.1002/sim.7654.

[36] Vergouwe Y, Royston P, Moons KGM, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical

Epidemiology. 2010 Feb;63(2):205-14. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0895435609001188.

[37] Wahl S, Boulesteix AL, Zierer A, Thorand B, van de Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Medical Research Methodology. 2016 Dec;16(1). Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0239-7.

[38] Framingham Coronary Heart Disease Risk Score;. Available from: https://www.mdcalc.com/framingham-coronary-heart-disease-risk-score.

[39] Seattle Heart Failure Model;. Available from: https://depts.washington.edu/shfm/index.php?width=1920&height=1080.

[40] QRISK3;. Available from: https://qrisk.org/three/.

[41] Marshall G, Warner B, MaWhinney S, Hammermeister K. Prospective prediction in the presence of missing data. Statistics in Medicine. 2002 Feb;21(4):561-70. Available from: http://doi.wiley.com/10.1002/sim.966.

[42] Janssen KJM, Vergouwe Y, Donders ART, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with Missing Predictor Values When Applying Clinical Prediction Models. Clinical Chemistry. 2009 May;55(5):994-1001. Available from: http://www.clinchem.org/cgi/doi/10.1373/clinchem.2008.115345.

[43] van Barreveld M, Hulleman M, Boersma LVA, Delnoy PPHM, Meine M, Tuinenburg AE, et al. Dutch outcome in implantable cardioverter-defibrillator therapy (DO-IT): registry design and baseline characteristics of a prospective observational cohort study to predict appropriate indication for implantable cardioverter-defibrillator. Netherlands Heart Journal. 2017 Oct;25(10):574-80. Available from: http://link.springer.com/10.1007/s12471-017-1016-x.

[44] Eilers PHC, Currie ID, Durbán M. Fast and compact smoothing on large multidimensional grids. Computational Statistics & Data Analysis. 2006 Jan;50(1):61-76. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167947304002270.

[45] Hofert M, Kojadinovic I, Maechler M, Yan J. copula: Multivariate Dependence with Copulas; 2018. Available from: http://cran.r-project.org/package=copula.

[46] Nelsen RB. Introduction to Copulas. New York, NY: Springer Science+Business Media, Inc; 2006. OCLC: 990566625.

[47] Honaker J, King G, Blackwell M. **Amelia** II: A Program for Missing Data. Journal of Statistical Software. 2011;45(7):1-47. Available from: http://www.jstatsoft.org/v45/i07/.

[48] White IR, Royston P. Imputing missing covariate values for the Cox model. Statistics in Medicine. 2009;28(15):1982-98.

[49] Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JA. Joint modelling rationale for chained equations. BMC Medical Research Methodology. 2014 Dec;14(1). Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-28.

[50] Liu J, Gelman A, Hill J, Su YS, Kropko J. On the stationary distribution of iterative imputations. Biometrika. 2014 Mar;101(1):155-73. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/ast044.

[51] Steyerberg EW. Clinical Prediction Models. Statistics for Biology and Health. New York, NY: Springer New York; 2009. Available from: http://link.springer.com/10.1007/978-0-387-77244-8.

[52] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002 Aug;21(15):2175-97. Available from: http://doi.wiley.com/10.1002/sim.1203.

[53] Crowther MJ, Lambert PC. A general framework for parametric survival analysis. Statistics in Medicine. 2014 Dec;33(30):5280-97. Available from: http://doi.wiley.com/10.1002/sim.6300.

[54] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187-220.

[55] Cox DR. Partial Likelihood. Biometrika. 1975 Aug;62(2):269. Available from: https://www.jstor.org/stable/2335362?origin=crossref.

[56] Lin DY. On the Breslow estimator. Lifetime Data Analysis. 2007 Dec;13(4):471-80. Available from: http://link.springer.com/10.1007/s10985-007-9048-y.

[57] Lambert P. STPM2: Stata module to estimate flexible parametric survival models; 2010. Published: Statistical Software Components, Boston College Department of Economics. Available from: https://ideas.repec.org/c/boc/bocode/s457128.html.

[58] StataCorp. Stata Statistical Software. College Station, TX: StataCorp LLC; 2021.

[59] Liu XR, Pawitan Y, Clements M. Parametric and penalized generalized survival models. Statistical Methods in Medical Research. 2018 May;27(5):1531-46. Available from: http://journals.sagepub.com/doi/10.1177/0962280216664760.

[60] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.

[61] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. No. 143 in Monographs on statistics and applied probability. Boca Raton: CRC Press, Taylor & Francis Group; 2015.

[62] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970 Feb;12(1):55-67. Available from: https://www.jstor.org/stable/1267351?origin=crossref.

[63] Hastie T. Ridge Regularization: An Essential Concept in Data Science. Technometrics. 2020 Oct;62(4):426-33. Available from: https://www.tandfonline.com/doi/full/10.1080/00401706.2020.1791959.

[64] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996:267-88.

[65] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010;33(1):1-22.

[66] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Wiley series in probability and statistics. Hoboken, N.J: J. Wiley; 2002.

[67] Harrell Jr FE. rms: Regression Modeling Strategies; 2017. Available from: https://CRAN.R-project.org/package=rms.

[68] Novomestky F. gaussquad: Collection of functions for Gaussian quadrature; 2013. R package version 1.0-2. Available from: https://CRAN.R-project.org/package=gaussquad.

[69] Bower H, Crowther MJ, Lambert PC. Strcs: A Command for Fitting Flexible Parametric Survival Models on the Log-hazard Scale. The Stata Journal: Promoting communications on statistics and Stata. 2016 Dec;16(4):989-1012. Available from: http://journals.sagepub.com/doi/10.1177/1536867X1601600410.

[70] Meier L, van de Geer S, Bühlmann P. The group lasso for logistic regression. Journal of the Royal Statistical Society Series B. 2008;70(1):53-71.

[71] Boyd SP, Vandenberghe L. Convex optimization. 18th ed. Cambridge: Cambridge Univ. Press; 2015. OCLC: 767754283.

[72] Fu A, Narasimhan B, Boyd S. **CVXR** : An *R* Package for Disciplined Convex Optimization. Journal of Statistical Software. 2020;94(14). Available from: http://www.jstatsoft.org/v94/i14/.

[73] Domahidi A, Chu E, Boyd S. ECOS: An SOCP solver for embedded systems. In: 2013 European Control Conference (ECC). Zurich: IEEE; 2013. p. 3071-6. Available from: https://ieeexplore.ieee.org/document/6669541/.

[74] Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Statistics in Medicine. 2013 Oct;32(23):4118-34. Available from: http://doi.wiley.com/10.1002/sim.5823.

[75] Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York; London: Springer; 2011. OCLC: 751583845.

[76] Austin PC, Harrell FE, Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine. 2020 Sep;39(21):2714-42. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8570.

[77] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine. 1999 Sep;18(17-18):2529-45. Available from: http://doi.wiley.com/10.1002/%28SICI%291097-0258%2819990915/30%2918%3A17/18%3C2529%3A%3AAID-SIM274%3E3.0.CO%3B2-5.

[78] Antolini L, Boracchi P, Biganzoli E. A time-dependent discrimination index for survival data. Statistics in Medicine. 2005 Dec;24(24):3927-44. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.2427.

[79] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996 Feb;15(4):361-87. Available from: http://doi.wiley.com/10.1002/%28SICI%291097-0258%2819960229%2915%3A4%3C361%3A%3AAID-SIM168%3E3.0.CO%3B2-4.

[80] Therneau T. _A Package for Survival Analysis in R_; 2021. Available from: https://CRAN.R-project.org/package=survival.

[81] Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, et al.. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models; 2021. Available from: https://cran.r-project.org/package=glmnet.

[82] Kook L, Hothorn T. Regularized Transformation Models: The tramnet Package. The R Journal. 2021;13(1):14.

[83] Wood SN. Generalized additive models: an introduction with R. 2nd ed. Chapman & Hall/CRC texts in statistical science. Boca Raton: CRC Press/Taylor & Francis Group; 2017.

[84] Fauvernier M, Roche L, Uhry Z, Tron L, Bossard N, Remontet L, et al. Multidimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2019 Nov;68(5):1233-57. Available from: https://onlinelibrary.wiley.com/doi/10.1111/rssc.12368.

[85] Fauvernier M, Remontet L, Uhry Z, Bossard N, Roche L. survPen: an R package for hazard and excess hazard modelling with multidimensional penalized splines. Journal of Open Source Software. 2019 Aug;4(40):1434. Available from: https://joss.theoj.org/papers/10.21105/joss.01434.

[86] Chouldechova A, Hastie T. Generalized additive model selection. arXiv preprint arXiv:150603850. 2015.

[87] Senn S. Statistical issues in drug development. 2nd ed. Chichester, England, Hoboken, NJ: John Wiley & Sons; 2007. OCLC: ocn180907943.

[88] Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ. 1995 Nov;311(7016):1356-9. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.311.7016.1356.

[89] Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018:k4245. Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.k4245.

[90] Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Annals of Internal Medicine. 2020 Jan;172(1):35. Available from: https://annals.org/aim/fullarticle/2755582/predictive-approaches-treatment-effect-heterogeneity-path-statement.

[91] Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International Journal of Epidemiology. 2016 Nov:dyw125. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw125.

[92] Murray EJ, Caniglia EC, Swanson SA, Hernández-Díaz S, Hernán MA. Patients and investigators prefer measures of absolute risk in subgroups for pragmatic randomized trials. Journal of Clinical Epidemiology. 2018 Nov;103:10-21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435617308922.

[93] Rubin DB. Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association. 2005 Mar;100(469):322-31. Available from: http://www.tandfonline.com/doi/abs/10.1198/016214504000001880.

[94] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974;66(5):688-701. Available from: http://content.apa.org/journals/edu/66/5/688.

[95] Hernan MA. A definition of causal effect for epidemiological research. Journal of Epidemiology & Community Health. 2004 Apr;58(4):265-71. Available from: https://jech.bmj.com/lookup/doi/10.1136/jech.2002.006361.

[96] Hernán MA, Hernández-Díaz S, Robins JM. Randomized Trials Analyzed as Observational Studies. Annals of Internal Medicine. 2013 Sep. Available from: http://annals.org/article.aspx?doi=10.7326/0003-4819-159-8-201310150-00709.

[97] Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. New England Journal of Medicine. 2017 Oct;377(14):1391-8. Available from: http://www.nejm.org/doi/10.1056/NEJMsm1605385.

[98] Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, "on behalf of" the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. Statistics in Medicine. 2020 Sep:1-27. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.8741.

[99] Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine. 1986 Sep;5(5):421-33. Available from: http://doi.wiley.com/10.1002/sim.4780050506.

[100] Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. Journal of Clinical Epidemiology. 2010 Jan;63(1):2-6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435608003168.

[101] Nelder JA, Wedderburn RWM. Generalized Linear Models. Journal of the Royal Statistical Society Series A (General). 1972;135(3):370-84.

[102] Agresti A. Foundations of linear and generalized linear models. Wiley series in probability and statistics. Hoboken, New Jersey: John Wiley & Sons Inc; 2015.

[103] Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar:m441. Available from: http://www.bmj.com/lookup/doi/10.1136/bmj.m441.

[104] van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research. 2018 Jul;(1-20):096228021878472. Available from: http://journals.sagepub.com/doi/10.1177/0962280218784726.

[105] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of statistics. 2004;32(2):407-99.

[106] Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Statistical Methods in Medical Research. 2020 May;29(11):3166-78. Available from: http://journals.sagepub.com/doi/10.1177/0962280220921415.

[107] Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, et al. Penalisation and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. Journal of Clinical Epidemiology. 2020 Dec;132:88-96. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435620312099.

[108] Lim M, Hastie T. Learning Interactions via Hierarchical Group-Lasso Regularization. Journal of Computational and Graphical Statistics. 2015 Jul;24(3):627-54. Available from: http://www.tandfonline.com/doi/full/10.1080/10618600.2014.938812.

[109] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalize learning problems. Statistics and Computing. 2015 Nov;25(6):1129-41. Available from: http://link.springer.com/10.1007/s11222-014-9498-5.

[110] Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. BMJ. 2015 Feb;350(feb19 2):h454. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.h454.

[111] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11(85):1-11.

[112] Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circulation: Cardiovascular Quality and Outcomes. 2014;7(1):163-9.

[113] Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association. 2018 Jul;113(523):1228-42. Available from: https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839.

[114] Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. Available from: http://link.springer.com/10.1023/A:1010933404324.

[115] Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. Journal of Computational and Graphical Statistics. 2018 Jan;27(1):209-19. Available from: https://www.tandfonline.com/doi/full/10.1080/10618600.2017.1356325.

[116] Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics. 2019 Apr;47(2). Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-47/issue-2/Generalized-random-forests/10.1214/18-AOS1709.full.

[117] Zeileis A, Hothorn T, Hornik K. Model-Based Recursive Partitioning. Journal of Computational and Graphical Statistics. 2008 Jun;17(2):492-514. Available from: http://www.tandfonline.com/doi/abs/10.1198/106186008X319331.

[118] Seibold H, Zeileis A, Hothorn T. Model-Based Recursive Partitioning for Subgroup Analyses. The International Journal of Biostatistics. 2016 May;12(1). Available from: http://www.degruyter.com/view/j/ijb.2016.12.issue-1/ijb-2015-0032/ijb-2015-0032.xml.

[119] Scarpa J, Bruzelius E, Doupe P, Le M, Faghmous J, Baum A. Assessment of Risk of Harm Associated With Intensive Blood Pressure Management Among Patients With Hypertension Who Smoke: A Secondary Analysis of the Systolic Blood Pressure Intervention Trial. JAMA Network Open. 2019 Mar;2(3):e190005. Available from: http://jamanetworkopen.jamanetwork.com/article.aspx?doi=10.1001/jamanetworkopen.2019.0005.

[120] Rigdon J, Baiocchi M, Basu S. Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. Trials. 2018 Dec;19(1):382. Available from: https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-018-2774-5.

[121] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology. 2014 Dec;14(1):137. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-137.

[122] Austin PC, Harrell FE, Steyerberg EW. Predictive performance of machine and statistical learning methods: Impact of data-generating processes on external validity in the "large N, small p" setting. Statistical Methods in Medical Research. 2021 Apr:096228022110028. Available from: http://journals.sagepub.com/doi/10.1177/09622802211002867.

[123] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006 Dec;25(24):4279-92. Available from: http://doi.wiley.com/10.1002/sim.2673.

[124] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019:1-29. Available from: http://doi.wiley.com/10.1002/sim.8086.

[125] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available from: https://www.R-project.org/.

[126] Le Saux N. A randomized, double-blind, placebo-controlled noninferiority trial of amoxicillin for clinically diagnosed acute otitis media in children 6 months to 5 years of age. Canadian Medical Association Journal. 2005 Feb;172(3):335-41. Available from: http://www.cmaj.ca/cgi/doi/10.1503/cmaj.1040771.

[127] Rovers MM, Glasziou P, Appelman CL, Burke P, McCormick DP, Damoiseaux RA, et al. Antibiotics for acute otitis media: a meta-analysis with individual patient data. The Lancet. 2006 Oct;368(9545):1429-35. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673606696062.

[128] International Stroke Trial Collaborative Group. The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. The Lancet. 1997 May;349(9065):1569-81. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673697040117.

[129] International Stroke Trial Collaborative Group, Sandercock PA, Niewada M, Członkowska A. The International Stroke Trial database. Trials. 2011 Dec;12(1). Available from: https://trialsjournal.biomedcentral.com/articles/10.1186/1745-6215-12-101.

[130] Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. European Journal of Epidemiology. 2020 Jul;35(7):615-7. Available from: http://link.springer.com/10.1007/s10654-020-00659-8.

[131] Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. Journal of Clinical Epidemiology. 2020 Sep;125:47-56. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0895435620300445.

[132] van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. Journal of Clinical Epidemiology. 2018;94:59-68. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0895435617303037.

[133] Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration. Annals of Internal Medicine. 2020 Jan;172(1):W1-W25. Available from: https://annals.org/aim/fullarticle/2755583/.

[134] Rekkas A, Paulus JK, Raman G, Wong JB, Steyerberg EW, Rijnbeek PR, et al. Predictive approaches to heterogeneous treatment effects: a scoping review. BMC Medical Research Methodology. 2020 Dec;20(1):264. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-01145-1.

[135] Lamont A, Lyons MD, Jaki T, Stuart E, Feaster DJ, Tharmaratnam K, et al. Identification of predicted individual treatment effects in randomized clinical trials. Statistical Methods in Medical Research. 2018;27(1):142-57. Available from: http://journals.sagepub.com/doi/10.1177/0962280215623981.

[136] Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011 Apr;12(2):270-82. Available from: https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxq060.

[137] Kang C, Janes H, Huang Y. Combining biomarkers to optimize patient treatment recommendations: Combine Markers for Treatment Selection. Biometrics. 2014 Sep;70(3):695-707. Available from: http://doi.wiley.com/10.1111/biom.12191.

[138] Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively Selecting a Target Population for a Future Comparative Study. Journal of the American Statistical Association. 2013 Jun;108(502):527-39. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.770705.

[139] Li J, Zhao L, Tian L, Cai T, Claggett B, Callegaro A, et al. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies: Predictive Enrichment. Biometrics. 2016 Sep;72(3):877-87. Available from: http://doi.wiley.com/10.1111/biom.12461.

[140] van Geloven N, Swanson S, Ramspek C, Luijken K, van Diepen M, Morris T, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol. 2020 Apr;35(7):619-30. ArXiv: 2004.06998. Available from: http://arxiv.org/abs/2004.06998.

[141] Hand DJ. Classifier Technology and the Illusion of Progress. Statistical Science. 2006 Feb;21(1):1-14. Available from: http://projecteuclid.org/euclid.ss/1149600839.

[142] Senn S. Statistical pitfalls of personalized medicine. Nature. 2018 Nov;563(7733):619-21. Available from: http://www.nature.com/articles/d41586-018-07535-2.

[143] Senn S. Cross-over trials in clinical research. 2nd ed. Statistics in practice. Chichester, England, New York: J. Wiley; 2002.

[144] Riley RD, Debray TPA, Fisher D, Hattle M, Marlin N, Hoogland J, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. Statistics in Medicine. 2020 Apr. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8516.

[145] Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. Statistics in Medicine. 2002 Oct;21(19):2909-16. Available from: http://doi.wiley.com/10.1002/sim.1295.

[146] Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, E Harrell Jr F, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Statistics in Medicine. 2021 Aug:sim.9154. Available from: https://onlinelibrary.wiley.com/doi/10.1002/sim.9154.

[147] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019 Dec;17(1):230. Available from: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1466-7.

[148] The IST-3 collaborative group. The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. The Lancet. 2012 Jun;379(9834):2352-63. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673612607685.

[149] Harrell FE, Califf RM. Evaluating the Yield of Medical Tests. JAMA. 1982 May;247(18):4.

[150] Bress AP, Greene T, Derington CG, Shen J, Xu Y, Zhang Y, et al. Patient Selection for Intensive Blood Pressure Management Based on Benefit and Adverse Events. Journal of the American College of Cardiology. 2021 Apr;77(16):1977-90. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0735109721005684.

[151] Olsen MK, Stechuchak KM, Oddone EZ, Damschroder LJ, Maciejewski ML. Which patients benefit most from completing health risk assessments: comparing methods to identify heterogeneity of treatment effects. Health Services and Outcomes Research Methodology. 2021 Feb. Available from: http://link.springer.com/10.1007/s10742-021-00243-x.

[152] Duan T, Rajpurkar P, Laird D, Ng AY, Basu S. Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy: A Machine Learning Experiment to Estimate Treatment Effects from Randomized Trial Data. Circulation: Cardiovascular Quality and Outcomes. 2019 Mar;12(3). Available from: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.005010.

[153] Rosenbaum PR. A Characterization of Optimal Designs for Observational Studies. Journal of the Royal Statistical Society: Series B (Methodological). 1991 Jul;53(3):597-610. Available from: http://doi.wiley.com/10.1111/j.2517-6161.1991.tb01848.x.

[154] Hansen BB. Full Matching in an Observational Study of Coaching for the SAT. Journal of the American Statistical Association. 2004 Sep;99(467):609-18. Available from: http://www.tandfonline.com/doi/abs/10.1198/016214504000000647.

[155] Colannino J, Damian M, Hurtado F, Langerman S, Meijer H, Ramaswami S, et al. Efficient Many-To-Many Point Matching in One Dimension. Graphs and Combinatorics. 2007 Jun;23(S1):169-78. Available from: http://link.springer.com/10.1007/s00373-007-0714-3.

[156] Hansen BB. The prognostic analogue of the propensity score. Biometrika. 2008 Feb;95(2):481-8. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asn004.

[157] Nguyen T, Debray TPA. The use of prognostic scores for causal inference with general treatment regimes. Statistics in Medicine. 2019 May;38(11):2013-29. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8084.

[158] van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings: A new Concordance Measure for External Validation of Risk Models. Statistics in Medicine. 2016 Oct;35(23):4136-52. Available from: http://doi.wiley.com/10.1002/sim.6997.

[159] Nieboer D, van der Ploeg T, Steyerberg EW. Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLOS ONE. 2016 Feb;11(2):e0148820. Available from: https://dx.plos.org/10.1371/journal.pone.0148820.

[160] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/.

[161] Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. Stroke. 2019 Dec;50(12). Available from: https://www.ahajournals.org/doi/10.1161/STR.0000000000000211.

[162] Sandercock, P. The third International Stroke Trial (IST-3), 2000-2015 [dataset]. University of Edinburgh & Edinburgh Clinical Trials Unit; 2016. Https://datashare.ed.ac.uk/handle/10283/1931; DOI: 10.7488/DS/1350. Available from: https://datashare.ed.ac.uk/handle/10283/1931.

[163] Agresti A. Categorical data analysis. 3rd ed. No. 792 in Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2013.

[164] Harrell F. Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine; 2018. Http://fharrell.com/post/hteview/. Available from: http://fharrell.com/post/hteview/.

[165] Efron B, Tibshirani R. Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association. 1997 Jun;92(438):548. Available from: https://www.jstor.org/stable/2965703?origin=crossref.

[166] Dekker ARJ, Verheij TJM, van der Velden AW. Inappropriate antibiotic prescription for respiratory tract indications: most prominent in adult patients. Family Practice. 2015 Apr:cmv019. Available from: https://academic.oup.com/fampra/article-lookup/doi/10.1093/fampra/cmv019.

[167] Fleming-Dutra KE, Hersh AL, Shapiro DJ, Bartoces M, Enns EA, File TM, et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. JAMA. 2016 May;315(17):1864. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.4151.

[168] Laxminarayan R, Duse A, Wattal C, Zaidi AKM, Wertheim HFL, Sumpradit N, et al. Antibiotic resistance—the need for global solutions. The Lancet Infectious Diseases. 2013 Dec;13(12):1057-98. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309913703189.

[169] Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. BMJ. 2010 Jun;340(may18 2):c2096-6. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.c2096.

[170] Tonkin-Crine S, Yardley L, Little P. Antibiotic prescribing for acute respiratory tract infections in primary care: a systematic review and meta-ethnography. Journal of Antimicrobial Chemotherapy. 2011 Oct;66(10):2215-23. Available from: https://academic.oup.com/jac/article-lookup/doi/10.1093/jac/dkr279.

[171] Lemiengre MB, van Driel ML, Merenstein D, Liira H, Mäkelä M, De Sutter AI. Antibiotics for acute rhinosinusitis in adults. Cochrane Database of Systematic Reviews. 2018 Sep;2018(9). Available from: https://doi.wiley.com/10.1002/14651858.CD006089.pub5.

[172] Young J. Antibiotics for adults with clinically diagnosed acute rhinosinusitis: a meta-analysis of individual patient data. Lancet. 2008;371:908-14.

[173] Kent DM, Hayward RA. Limitations of Applying Summary Results of Clinical Trials to Individual Patients: The Need for Risk Stratification. JAMA. 2007 Sep;298(10):1209. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.298.10.1209.

[174] Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Medical Research Methodology. 2006 Dec;6(1):18. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-6-18.

[175] Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. International Journal of Epidemiology. 2016 Jul:dyw118. Available from: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dyw118.

[176] Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010 Aug;340(feb05 1):c221-1. Available from: http://www.bmj.com/cgi/doi/10.1136/bmj.c221.

[177] Debray TPA, Riley RD, Rovers MM, Reitsma JB, Moons KGM, Cochrane IPD Meta-analysis Methods group. Individual Participant Data (IPD) Meta-analyses of Diagnostic and Prognostic Modeling Studies: Guidance on Their Use. PLOS Medicine. 2015 Oct;12(10):e1001886. Available from: http://dx.plos.org/10.1371/journal.pmed.1001886.

[178] Venekamp RP, Hoogland J, van Smeden M, Rovers MM, De Sutter AI, Merenstein D, et al. Identifying adults with acute rhinosinusitis in primary care that benefit most from antibiotics: protocol of an individual patient data meta-analysis using multivariable risk prediction modelling. BMJ Open. 2021 Jul;11(7):e047186. Available from: https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2020-047186.

[179] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Annals of Internal Medicine. 2015 Jan;162(1):55-63. Available from: https://www.acpjournals.org/doi/10.7326/M14-0697.

[180] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of Internal Medicine. 2015 Jan;162(1):W1-W73. Available from: https://www.acpjournals.org/doi/10.7326/M14-0698.

[181] Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data: The PRISMA-IPD Statement. JAMA. 2015 Apr;313(16):1657. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2015.3656.

[182] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug:l4898. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.l4898.

[183] Steyerberg EW, Nieboer D, Debray TPA, Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Statistics in Medicine. 2019 Sep;38(22):4290-309. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8296.

[184] Meinshausen N. Relaxed Lasso. Computational Statistics & Data Analysis. 2007 Sep;52(1):374-93. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167947306004956.

[185] Hastie T, Narasimhan B, Tibshirani R. The Relaxed Lasso. CRAN; 2021. Available from: https://cran.r-project.org/web/packages/glmnet/vignettes/relax.pdf.

[186] Thao LTP, Geskus R. A comparison of model selection methods for prediction in the presence of multiply imputed data. Biometrical Journal. 2019 Mar;61(2):343-56. Available from: http://doi.wiley.com/10.1002/bimj.201700232.

[187] IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology. 2014 Dec;14(1):25. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-14-25.

[188] Stalman W, Melker RAD. The end of antibiotic treatment in adults with acute sinusitis-like complaints in general practice? A placebo-controlled double-blind randomized doxycycline trial. British Journal of General Practice. 1997:6.

[189] Kaiser L, Morabia A, Stalder H, Ricchetti A, Auckenthaler R, Terrier F, et al. Role of Nasopharyngeal Culture in Antibiotic Prescription for Patients with Common Cold or Acute Sinusitis. European Journal of Clinical Microbiology and Infections Diseases. 2001 Jul;20(7):0445-51. Available from: http://link.springer.com/10.1007/s100960100544.

[190] De Sutter AI, De Meyere MJ, Christiaens TC, Van Driel ML, Peersman W, De Maeseneer JM. Does amoxicillin improve outcomes in patients with purulent rhinorrhea? A pragmatic randomized double-blind controlled trial in family practice. The Journal of Family Practice. 2002 Apr;51(4):317-23.

[191] Bucher HC. Effect of Amoxicillin-Clavulanate in Clinically Diagnosed Acute Rhinosinusitis: A Placebo-Controlled, Double-blind, Randomized Trial in General Practice. Archives of Internal Medicine. 2003 Aug;163(15):1793. Available from: http://archinte.jamanetwork.com/article.aspx?doi=10.1001/archinte.163.15.1793.

[192] Varonen H, Kunnamo I, Savolainen S, Mäkelä M, Revonta M, Ruotsalainen J, et al. Treatment of acute rhinosinusitis diagnosed by clinical criteria or ultrasound in primary care. Scandinavian Journal of Primary Health Care. 2003 Jan;21(2):121-6. Available from: http://www.tandfonline.com/doi/full/10.1080/02813430310001743.

[193] Meltzer E, Bachert C, Staudinger H. Treating acute rhinosinusitis: Comparing efficacy and safety of mometasone furoate nasal spray, amoxicillin, and placebo. Journal of Allergy and Clinical Immunology. 2005 Dec;116(6):1289-95. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0091674905019342.

[194] Merenstein D, Whittaker C, Chadwell T, Wegner B, D'Amico F. Are antibiotics beneficial for patients with sinusitis complaints? A randomized double-blind clinical trial. The Journal of Family Practice. 2005 Feb;54(2):144-51.

[195] Williamson IG, Rumsby K, Benge S, Moore M, Smith PW, Cross M, et al. Antibiotics and Topical Nasal Steroid for Treatment of Acute Maxillary Sinusitis: A Randomized Controlled Trial. JAMA. 2007 Dec;298(21):2487. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.298.21.2487.

[196] Schering-Plough Research Institute. Efficacy and Safety of 200 mcg QD or 200 mcg BID mometasone fuorate (MFNS) vs amoxicillin vs placebo as primary treatment of subjects with acute rhinosinusitis (protocol P02692). Kenilworth: Schering-Plough Research Institute,; 2003.

[197] Garbutt JM, Banister C, Spitznagel E, Piccirillo JF. Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial. JAMA. 2012 Feb;307(7):685. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2012.138.

[198] Takada T, Hoogland J, Hansen JG, Lindbaek M, Autio T, Alho OP, et al. Diagnostic prediction models for computed tomography-confirmed acute rhinosinusitis and culture-confirmed acute bacterial rhinosinusitis in adults presenting to primary care: an individual participant data meta-analysis. British Journal of General Practice. 2022 May:BJGP.2021.0585. Available from: http://bjgp.org/lookup/doi/10.3399/BJGP.2021.0585.

[199] Autio TJ, Koskenkorva T, Koivunen P, Alho OP. Inflammatory Biomarkers During Bacterial Acute Rhinosinusitis. Current Allergy and Asthma Reports. 2018 Feb;18(2):13. Available from: http://link.springer.com/10.1007/s11882-018-0761-2.

[200] Glinz D, Georg Hansen J, Trutmann C, Schaller B, Vogt J, Diermayr C, et al. Single-use device endoscopy for the diagnosis of acute bacterial rhinosinusitis in primary care: A pilot and feasibility study. Clinical Otolaryngology. 2021 Sep;46(5):1050-6. Available from: https://onlinelibrary.wiley.com/doi/10.1111/coa.13785.

[201] Carlton HC, Savović J, Dawson S, Mitchelmore PJ, Elwenspoek MMC. Novel point-of-care biomarker combination tests to differentiate acute bacterial from viral respiratory tract infections to guide antibiotic prescribing: a systematic review. Clinical Microbiology and Infection. 2021 Aug;27(8):1096-108. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1198743X21002573.

[202] Ross M, Henao R, Burke TW, Ko ER, McClain MT, Ginsburg GS, et al. A comparison of host response strategies to distinguish bacterial and viral infection. PLOS ONE. 2021 Dec;16(12):e0261385. Available from: https://dx.plos.org/10.1371/journal.pone.0261385.

[203] Sweeney TE, Wong HR, Khatri P. Robust classification of bacterial and viral infections via integrated host gene expression diagnostics. Science Translational Medicine. 2016 Jul;8(346). Available from: https://www.science.org/doi/10.1126/scitranslmed.aaf7165.

[204] quoteresearch. Quote Investigator; 2013. Available from: https://quoteinvestigator.com/2013/10/20/no-predict/.

[205] Breiman L. Statistical Modeling: The Two Cultures. Statist Sci. 2001;16(3):199-231.

[206] Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagnostic and Prognostic Research. 2021 Dec;5(1):1. Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-020-00090-3.

[207] Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM; 2020. p. 1-12. Available from: https://dl.acm.org/doi/10.1145/3313831.3376718.

[208] Imbens G, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015.

[209] Bühlmann P. Causal statistical inference in high dimensions. Mathematical Methods of Operations Research. 2013 Jun;77(3):357-70. Available from: http://link.springer.com/10.1007/s00186-012-0404-7.

[210] Westreich D, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. American Journal of Epidemiology. 2013 Feb;177(4):292-8. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kws412.

[211] Sperrin M, Jenkins D, Martin GP, Peek N. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. Journal of the American Medical Informatics Association. 2019 Dec;26(12):1675-6. Available from: https://academic.oup.com/jamia/article/26/12/1675/5625126.

[212] Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences. 2016 Jul;113(27):7345-52. Available from: https://pnas.org/doi/full/10.1073/pnas.1510507113.

Appendices

Summary

Recent decades have seen many new developments in prediction modeling, including the rise of algorithmic or machine learning methods and a rapid increase in the amount of data and the types of data that are becoming available. In this respect, there is a clear synergy between developments in the fields of statistics and computer science. At the same time, the sharing and interchanging of ideas and methods that have been developed in neighboring research fields in increasingly common. For instance, the causal inference frameworks that have classically been used for the estimation of population averaged effects prove fruitful when interest is in causal prediction. Likewise, there is a synergy between developments in methods for high-dimensional data and for low-dimensional data. For example, regularization methods are essential for high-dimensional data to obtain any solution, stimulating their further development, and yet also continue to provide useful results for low-dimensional settings. Concurrently, there has been a rapid increase in the ease with which complex models can be used in medical practice, with easily available online tools allowing for detailed input of covariate data and behind the scene handling of required transformations and the handling of missing data. In light of these developments, this dissertation focuses on recent advances in low-dimensional clinical prediction modeling in the medical domain, with a special interest in the handling of missing data, regularization methods, and the causal prediction of possibly heterogeneous treatment effect.

After a general introduction in **Chapter 1**, **Chapter 2** handles challenges presented by missing data during clinical prediction model development and real-world application. While these challenges have received considerable attention in the development setting, there is only sparse research on the handling of missing data in applied settings. The main unique feature of handling missing data in these settings is that missing data methods have to be performed for a single new individual, precluding direct application of mainstay methods used during model development. A seemingly straightforward, but often neglected, consequence of the use of missing data methods in clinical practice, is that these methods should also be part of the validation process. This chapter compares existing and new methods to account for missing data for a new individual in the context of prediction. These methods are based on (i) submodels based on observed data only, (ii) marginalization over the missing variables, or (iii) imputation based on fully conditional specification (also known as chained equations). They were compared in an internal validation setting to highlight

the use of missing data methods that transfer to practice while validating a model. As a reference, they were compared to the use of multiple imputation by chained equations in a set of test patients, because this has been used in validation studies in the past. The methods were evaluated in a simulation study where performance was measured by means of optimism corrected C-statistic and mean squared prediction error. Furthermore, they were applied in data from a large Dutch cohort of prophylactic implantable cardioverter defibrillator patients.

In the context of time-to-event data, **Chapter 3** proposes to combine the benefits of flexible parametric survival modeling and regularization to improve risk prediction modeling. Thereto, ridge, lasso, elastic net, and group lasso penalties were combined with both log hazard and log cumulative hazard models. The log (cumulative) hazard in these models is represented by a flexible function of time that may depend on the covariates (*i.e.*, covariate effects may be time-varying). It is shown that the optimization problem for the proposed models can be formulated as a convex optimization problem. A user-friendly `R` implementation is provided for model fitting and penalty parameter selection based on cross-validation in `R` package **regsurv**. Simulation study results show the advantage of regularization in terms of increased out-of-sample prediction accuracy and improved calibration and discrimination of predicted survival probabilities, especially when sample size was relatively small with respect to model complexity. An applied example illustrates the proposed methods. In summary, this chapter provides the required theoretical developments and an easily accessible implementation of regularized parametric survival modeling, and suggest that it improves out-of-sample prediction performance.

Chapter 4 focuses on the prediction of individual treatment effect from randomized clinical trials. Randomized trials typically estimate average relative treatment effects, but decisions on the benefit of a treatment are possibly better informed by more individualized predictions of the absolute treatment effect. In case of a binary outcome, these predictions of absolute individualized treatment effect require knowledge of the individual's risk without treatment and incorporation of a possibly differential treatment effect (*i.e.*, varying with patient characteristics). This chapter lays out the causal structure of individualized treatment effect in terms of potential outcomes and describe the required assumptions that underlie a causal interpretation of its prediction. Subsequently, regression models and model estimation techniques are described that can be used to move from average to more individualized treatment effect predictions. The main focus is on logistic regression-based methods that are both well-known

and naturally provide the required probabilistic estimates. Key components from both causal inference and prediction research combine to arrive at individualized treatment effect predictions. While the separate components are well-known, their successful incorporation is an ongoing field of research. This chapter cuts the problem down to its essentials in the setting of a randomized trial, discusses the importance of a clear definition of the estimand of interest, provides insight into the required assumptions, and gives guidance with respect to modeling and estimation options. Simulated data illustrate the potential of different modeling options across scenarios that vary both average treatment effect and treatment effect heterogeneity. Two applied examples illustrate individualized treatment effect prediction in randomized trial data.

As a continuation of the work described in **Chapter 4**, **Chapter 5** describes existing and novel methodology to evaluate individualized treatment effect models in terms of discrimination and calibration performance. Models for this purpose are increasingly common in the published literature. Aiming to facilitate the validation of prediction models for individualized treatment effects, the classical concepts of discrimination and calibration are as used in regular (associative) prediction are extended to the class of causal prediction models. Working within the potential outcomes framework, the statistical properties of existing statistics (including the c-for-benefit) are described. Subsequently novel model-based alternative are proposed. The main focus is on randomized trials with binary endpoints. Simulated data provide insight into the characteristics of discrimination and calibration statistics, and further illustrate all methods in a trial in acute ischemic stroke treatment. Results demonstrate that the proposed model-based statistics had the best characteristics in terms of bias and variance. While resampling methods to adjust for optimism of performance estimates in the development data were effective on average, they had a high variance across replications, limiting their accuracy in any particular applied analysis. Thereto, individualized treatment effect models are best validated in external data rather than in the original development sample.

Lastly, **Chapter 6** presents an applied study analyzing possible treatment effect heterogeneity in the antibiotic treatment of adults with clinically diagnosed acute rhinosinusitis (ARS). In line with the developments in **Chapter 4**, this chapters describes the joint prediction of prognosis with and without antibiotic treatment in terms of cure rate at 8-15 days. Nine double-blind placebo-controlled trials (n=2,539) were available for this task, with concomitant challenges in terms of systematically missing data and the clustering of data within individual studies. Congenial methods for imputation in mixed modeling con-

text were used prior to the subsequent development of a logistic mixed effects model to predict the probability of being cured at 8-15 days. Predictors included individual level demographic characteristics, common signs and symptoms, antibiotic treatment assignment, and study-level characteristics. Internal-external cross-validation was used to estimate out-of-sample prediction performance. In conclusion, a prediction model based on the measurements available in this IPD-MA did not provide sufficient performance to adequately predict prognosis or antibiotic treatment benefit in adults presenting to primary care with clinically diagnosed ARS.

To conclude, this dissertation has provided methods for the handling of missing data during prediction model validation and prediction model application in practice, has provided a flexible integration of parametric survival modeling and several regularization techniques, and has explored the combination classical prediction modeling methods and causal inference methods for the purpose of individualized treatment effect prediction. Free and open-source `R` code is available for all of these endeavors. Together, the developments and guidance in this dissertation may aid in the process of prediction model development, validation, and implementation in practice.

Samenvatting

De afgelopen decennia zijn er veel nieuwe ontwikkelingen geweest op het gebied van predictiemodellen, waaronder de opkomst van algoritmische of machine learning-methoden, en een snelle toename van de hoeveelheid en soorten van data die beschikbaar komen. Er is wat dat betreft een duidelijke synergie tussen ontwikkelingen op het gebied van statistiek en informatica. Tegelijkertijd komt het delen en uitwisselen van ideeën en methoden uit aangrenzende onderzoeksgebieden steeds vaker voor. Als voorbeeld blijken veel aspecten van methoden die klassiek gebruikt worden voor causale inferentie over (populatie) gemiddelde effecten ook bruikbaar voor meer geïndividualiseerde causale predicties. Ook is er een synergie tussen ontwikkelingen in methoden voor hooge laag-dimensionale data. Bij hoogdimensionale data zijn regularisatiemethoden essentieel om überhaupt een oplossing te vinden, waardoor hun verdere ontwikkeling wordt gestimuleerd. Dit levert weer belangrijk bijdragen in de laagdimensionale context waar veelal de eerste methodologische stappen werden gezet. Daarnaast is het gemak waarmee complexe modellen in de medische praktijk kunnen worden gebruikt snel toegenomen. Gemakkelijk beschikbare online tools maken gedetailleerde invoer van data eenvoudig, terwijl achter de schermen de vereiste transformaties en de behandeling van ontbrekende data afgehandeld worden. In het licht van deze ontwikkelingen richt dit proefschrift zich op recente ontwikkelingen in laagdimensionale klinische predictiemodellen in het medische domein, en in het bijzonder op het omgaan met ontbrekende data, regularisatiemethoden, en de causale predictie van mogelijk heterogene behandeling effecten.

Na een algemene inleiding in Hoofdstuk 1, behandelt Hoofdstuk 2 uitdagingen die ontstaan door ontbrekende data tijdens de ontwikkeling en vooral toepassing van klinische predictiemodellen in de praktijk. Hoewel deze uitdagingen veel aandacht hebben gekregen in de context van modelontwikkeling, is er weinig onderzoek gedaan naar het moment van toepassing. Het belangrijkste kenmerk van deze laatste situatie is dat methoden voor ontbrekende data moeten worden uitgevoerd voor een enkel nieuw individu, waardoor directe toepassing van veel gangbare methoden niet mogelijk is. Een ogenschijnlijk voor de hand liggende, maar vaak verwaarloosde consequentie van het gebruik van methoden voor ontbrekende data in de praktijk, is dat deze methoden ook deel moeten uitmaken van het validatieproces. Hoofdstuk 2 vergelijkt bestaande en nieuwe methoden om rekening te houden met ontbrekende data voor een nieuw individu in de context van predictie. Deze methoden zijn gebaseerd op (i) submodellen op basis

van enkel de geobserveerde data, (ii) marginalisering over de ontbrekende data, of (iii) imputatie op basis van full conditional specification (ook wel bekend als chained equations). Ze worden vergeleken in een interne validatie setting om te benadrukken dat methoden voor ontbrekende data die bedoeld zijn voor de praktijk ook meegenomen moeten worden in de validatie. Ter referentie werden ze vergeleken met het gebruik van meervoudige imputatie door chained equations bij een reeks testpatiënten, omdat dit in het verleden in validatiestudies is gebruikt. De methoden werden geëvalueerd in een simulatiestudie waarbij methoden werden vergeleken op basis van een voor optimisme gecorrigeerde C-statistiek en gemiddelde kwadratische predictiefout. Bovendien werden ze toegepast bij de ontwikkeling van een predictiemodel op basis van een groot Nederlands cohort van patiënten met een defibrillator.

In de context van time-to-event data stelt Hoofdstuk 3 voor om de voordelen van flexibele parametrische survival modellen en regularisatie te combineren om risicopredictie te verbeteren. Daartoe werden ridge, lasso, elastic net, en groep-slasso penalty's gecombineerd met zowel log-hazard als log-cumulatieve-hazard modellen. De log (cumulatieve) hazard in deze modellen wordt weergegeven door een flexibele functie van de tijd die kan afhangen van de covariaten; d.w.z. covariaat effecten kunnen in de tijd variëren. Er wordt aangetoond dat het optimalisatieprobleem voor de voorgestelde modellen kan worden geformuleerd als een convex optimalisatieprobleem. Ook is er in R package regsurv een gebruiksvriendelijke software implementatie gemaakt voor het optimaliseren van de modellen en voor de selectie van penalty parameters op basis van kruisvalidatie. Resultaten van simulatiestudies tonen het voordeel van regularisatie in termen van predictie-accuratesse en verbeterde calibratie en discriminatie van de voorspelde overlevingskansen in nieuwe data. Dit was vooral te zien wanneer de steekproefomvang relatief klein was met betrekking tot de modelcomplexiteit. Een toegepast voorbeeld illustreert de voorgestelde methoden. Samenvattend biedt dit hoofdstuk de vereiste theoretische ontwikkelingen en een toegankelijke implementatie van geregulariseerde parametrische survival modellen, en suggereert het dat het dat de predicties in nieuwe data verbetert.

Hoofdstuk 4 richt zich op de predictie van individuele behandeleffecten op basis van gerandomiseerde klinische studies. Gerandomiseerde onderzoeken schatten doorgaans gemiddelde relatieve behandeleffecten, maar beslissingen over het voordeel van een behandeling worden in de praktijk mogelijk beter geïnformeerd door meer geïndividualiseerde schattingen van het absolute behandeleffect. In het geval van een binaire uitkomst vereisen deze meer geïndividualiseerde predicties van behandelingseffect zowel kennis van het individuele risico zonder

behandeling als van het mogelijk heterogene relatieve effect van de behandeling. Dit hoofdstuk legt de causale structuur van geïndividualiseerd behandeleffect uit in termen van potential outcomes en beschrijft de vereiste aannames die ten grondslag liggen aan een causale interpretatie van de predictie ervan. Vervolgens worden regressiemodellen en schattingstechnieken beschreven die kunnen worden gebruikt om van gemiddelde naar meer geïndividualiseerde behandeleffectpredicties te komen. De focus ligt hierbij op logistische regressietechnieken die algemeen bekend zijn en van nature de vereiste probabilistische schattingen geven. Hierbij worden sleutelcomponenten van zowel causale inferentie- als predictiesonderzoek gecombineerd om te komen tot geïndividualiseerde predicties van behandeleffecten. Hoewel de afzonderlijke componenten bekend zijn, is hun integratie een actueel onderzoeksgebied. In dit hoofdstuk wordt het probleem teruggebracht tot de essentie in de context van een gerandomiseerde trial, wordt ingegaan op het belang van een duidelijke definitie van te schatten grootheid (estimand), wordt inzicht gegeven in de benodigde aannames, en worden adviezen gegeven met betrekking tot modellerings- en schattingsmogelijkheden. Gesimuleerde data illustreren de eigenschappen van de verschillende opties in scenario's die zowel het gemiddelde behandelingseffect als de heterogeniteit van het behandelingseffect variëren. Twee toegepaste voorbeelden illustreren de geïndividualiseerde predictie van behandeleffecten in gerandomiseerde onderzoeksdata.

Als voortzetting van het werk beschreven in Hoofdstuk 4, beschrijft Hoofdstuk 5 bestaande en nieuwe methoden om modellen voor geïndividualiseerde predicties van behandeleffect te evalueren in termen van discriminatie en calibratie. Zulke modellen komen steeds vaker voor in de literatuur. Om de validatie van predictiemodellen voor geïndividualiseerde behandeleffecten te vergemakkelijken, worden de klassieke concepten van discriminatie en calibratie, zoals gebruikt in reguliere (associatieve) predictie, uitgebreid naar de klasse van causale predictiemodellen. Werkend binnen het potential outcomes raamwerk worden de statistische eigenschappen van bestaande statistieken, zoals de c-for-benefit, beschreven. Vervolgens worden nieuwe, op modellen gebaseerde alternatieven voorgesteld. De focus ligt daarbij op gerandomiseerde studies met binaire eindpunten. Gesimuleerde data geven inzicht in de kenmerken van de discriminatie- en calibratiestatistieken, en alle methoden worden geïllustreerd in een studie naar de behandeling van acute herseninfarcten. De resultaten tonen dat de voorgestelde maten de beste kenmerken hadden in termen van bias en variantie. Hoewel resampling-methoden om te corrigeren voor optimisme tijdens interne validatie gemiddeld effectief waren, hadden ze een grote variantie tussen repli-

caties. Dit limiteert hun waarde in een specifieke toepassing. Daartoe kunnen modellen voor de predictie van geïndividualiseerde behandeleffecten het best worden gevalideerd in externe data in plaats van in de oorspronkelijke ontwikkelingssteekproef.

Ten slotte presenteert Hoofdstuk 6 een toegepaste studie, waarin de mogelijke heterogeniteit van het effect van antibioticabehandeling wordt geanalyseerd bij volwassenen met klinisch gediagnosticeerde acute rhinosinusitis (ARS). In overeenstemming met de ontwikkelingen in Hoofdstuk 4, beschrijft dit hoofdstuk de gezamenlijke predictie van de prognose met en zonder antibioticabehandeling in termen van het genezingspercentage na 8-15 dagen. Er waren negen dubbelblinde, placebogecontroleerde onderzoeken (n=2.539) beschikbaar, waarbij rekening gehouden moest worden systematisch ontbrekende data en de clustering van data binnen individuele onderzoeken. Bij imputatie werd rekening gehouden met de uiteindelijk mixed effects structuur van het logistische predictiemodel. Voorspellers waren onder meer demografische kenmerken op individueel niveau, algemene tekenen en symptomen, toewijzing van antibioticabehandeling en kenmerken op studieniveau. Interne-externe kruisvalidatie werd gebruikt om de predictiesprestaties voor nieuwe data te schatten. Samengevat, waren de predicties van zowel prognose als behandeleffect van onvoldoende kwaliteit om de beslissing aangaande antibioticabehandeling bij volwassenen met ARS te ondersteunen.

Concluderend heeft dit proefschrift methoden opgeleverd voor het omgaan met ontbrekende data tijdens de validatie van een predictiemodel en de toepassing van een predictiemodel in de praktijk, maakt het flexibele geregulariseerde parametrische survival modellen beschikbaar, en heeft het de combinatie van klassieke predictie methoden en causale methoden onderzocht ten behoeve van geïndividualiseerde predicties van behandeleffecten. De open-source R-code is vrijelijk beschikbaar. Samen kunnen de ontwikkelingen en praktische handvatten in dit proefschrift helpen bij het proces van ontwikkeling, validatie en implementatie van predictiemodellen in de praktijk.

List of research output

Publications included in this dissertation

- Hoogland J, Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*. 2020;39(25):3591-3607.
- Hoogland J, Debray TPA, Crowther MJ, Riley RD, IntHout J, Reitsma JB, Zwinderman AH. Regularized parametric survival modeling to improve risk prediction models. (*under revision*)
- Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, E. Harrell Jr F, et al. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*. 2021;40(26):5961-5981.
- Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment. *arXiv preprint arXiv:2209.06101*. (*under revision*)
- Hoogland J, Takada T, Smeden M, Rovers MM, de Sutter AI, Merenstein D, van Essen GA, Kaiser L, Liira H, Little P, Bucher HC, Moons KGM, Reitsma JB, Venekamp RP. Prognosis and prediction of antibiotic benefit in adults with clinically diagnosed acute rhinosinusitis: an individual participant data meta-analysis. (*under revision*)

Publications not included in this dissertation

- van de Wiel MA, Leday GGR, Hoogland J, Heymans MW, van Zwet EW, Zwinderman AH. Think before you shrink: Alternatives to default shrinkage methods can improve prediction accuracy, calibration and coverage. *arXiv preprint arXiv:2301.09890*.
- Efthimiou O, Hoogland J, Debray TPA, Seo M, Furukawa TA, Egger M, et al. Measuring the performance of prediction models to personalize treatment choice. *Statistics in Medicine*. 2023;42(8):1188–206.
- Belias M, Rovers MM, Hoogland J, Reitsma JB, Debray TPA, IntHout J. Predicting personalised absolute treatment effects in individual partic-

ipant data meta-analysis: An introduction to splines. *Research Synthesis Methods*. 2022;13(2):255–83.

- Leeuwenberg AM, Reitsma JB, Van den Bosch LGLJ, Hoogland J, van der Schaaf A, Hoebers FJP, et al. The relation between prediction model performance measures and patient selection outcomes for proton therapy in head and neck cancer. *Radiotherapy and Oncology*. 2023;179:109449.

- Takada T, Hoogland J, van Lieshout C, Schuit E, Collins GS, Moons KGM, et al. Accuracy of approximations to recover incompletely reported logistic regression models depended on other available information. *Journal of Clinical Epidemiology*. 2022;143:81–90.

- Takada T, Hoogland J, Hansen JG, Lindbaek M, Autio T, Alho OP, et al. Diagnostic prediction models for computed tomography-confirmed acute rhinosinusitis and culture-confirmed acute bacterial rhinosinusitis in adults presenting to primary care: an individual participant data meta-analysis. *Br J Gen Pract*. 2022;BJGP.2021.0585.

- Venekamp RP, Hoogland J, van Smeden M, Rovers MM, De Sutter AI, Merenstein D, et al. Identifying adults with acute rhinosinusitis in primary care that benefit most from antibiotics: protocol of an individual patient data meta-analysis using multivariable risk prediction modelling. *BMJ Open*. 2021;11(7):e047186.

- Nijman SWJ, Groenhof TKJ, Hoogland J, Bots ML, Brandjes M, Jacobs JJL, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. *Journal of Clinical Epidemiology*. 2021;134:22–34.

- Nijman SWJ, Hoogland J, Groenhof TKJ, Brandjes M, Jacobs JJL, Bots ML, et al. Real-time imputation of missing predictor values in clinical practice. *European Heart Journal - Digital Health*. 2021;2(1):154–64.

- Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;m1328.

- Takada T, Hoogland J, Yano T, Fujii K, Fujiishi R, Miyashita J, et al. Added value of inflammatory markers to vital signs to predict mortality in patients suspected of severe infection. *The American Journal of Emergency Medicine*. 2020;38(7):1389–95.

- Riley RD, Debray TPA, Fisher D, Hattle M, Marlin N, Hoogland J, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. *Statistics in Medicine*. 2020;39(15):2115-2137
- Wijn SRW, Rovers MM, Le LH, Belias M, Hoogland J, IntHout J, et al. Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: a scoping review. *BMJ Open*. 2019;9(8):e028751.
- Hoogland J, Post B, de Bie RMA. Overall and Disease Related Mortality in Parkinson's Disease – a Longitudinal Cohort Study. *Journal of Parkinson's Disease*. 2019;9(4):767–74.
- Hoogland J, Boel JA, Bie RMA, Schmand BA, Geskus RB, Dalrymple-Alford JC, et al. Risk of Parkinson's disease dementia related to level I MDS PD-MCI. *Movement Disorders*. 2019;34(3):430–5.
- Hoogland J, van Wanrooij LL, Boel JA, Goldman JG, Stebbins GT, Dalrymple-Alford JC, et al. Detecting Mild Cognitive Deficits in Parkinson's Disease: Comparison of Neuropsychological Tests: Detecting Mild Cognitive deficits in PD. *Movement Disorders*. 2018;33(11):1750–9.
- Hoogland J, Boel JA, de Bie RMA, Geskus RB, Schmand BA, Dalrymple-Alford JC, et al. Mild cognitive impairment as a risk factor for Parkinson's disease dementia: MCI as a Risk Factor For PDD. *Movement Disorders*. 2017;32(7):1056–65.
- Geurtsen Gert J, Hoogland Jeroen, Goldman Jennifer G, Schmand Ben A, Troster Alexander I, Burn David J, et al. Parkinson's disease mild cognitive impairment: application and validation of the criteria. *Journal of Parkinson's Disease*. 2014;(2):131–7.
- Hoogland J, de Bie RMA, Williams-Gray CH, Muslimović D, Schmand B, Post B. Catechol-O-methyltransferase val158met and cognitive function in Parkinson's disease: COMT val158met and Cognitive Function in PD. *Movement Disorders*. 2010;25(15):2550–4.

Software contributions

- R-package **regsurv** `https://github.com/jeroenhoogland/regsurv`

Curriculum vitae

Jeroen Hoogland was born on the 1st of December, 1983, in Alphen a/d Rijn, the Netherlands. He completed his secondary education at Lyceum Sancta Maria in Haarlem, and obtained a M.Sc. degree in psychology (University of Amsterdam), a M.Sc degree in medicine and training as a medical doctor (University of Amsterdam), and a M.Sc. degree in biostatistics (Hasselt University). In the meantime, he wrote several papers in the field of neurology and worked as a postgraduate house officer (ANIOS) at the Neurology department of the Onze Lieve Vrouwe Gasthuis in Amsterdam.

In 2018 he started working on the present dissertation at UMC Utrecht, under promotors prof. dr. K.G.M. Moons and prof. dr. M.M. Rovers, co-promotors dr. T.P.A. Debray and dr. J. in 't Hout, and dr. J.B Reitsma. Key research topics revolved around prediction modeling methods and included missing data, survival analysis, individualized treatment effect prediction, and the joint analyses of multiple data sets. He presented his work at several biostatistics conferences. Next to research, he contribute to courses on missing data, survival analysis and mixed effects model.

In 2022, he joined the Epidemiology and Data Science department at the Amsterdam UMC to become an assistant professor, where he currently combines research activities, statistical consultancy, and teaching. He is also a member of the board of BMS-ANed, which jointly represents the biometrical section of the Netherlands Society for Statistics and Operations Research (BMS) and the Dutch region of the International Biometric Society (ANed).

Acknowledgements

This dissertation was supported by many people in many ways, and could not have been completed without them.

First of all, I would like to thank my promotors prof. dr. K.G.M. Moons and prof. dr. M.M. Rovers, co-promotors dr. T.P.A. Debray and dr. J. in 't Hout, and dr. J.B Reitsma for providing me with the opportunity to work on this project under their supervision. Each of you has guided me with your own unique blend of skills and it has been a privilege to work with and learn from you.

Furthermore, I would like to thank the co-authors that contributed to the chapters of this dissertation: prof. dr. A.H. Zwinderman, prof. dr. R.D. Riley, prof. dr. F.E. Harrell Jr, dr. M.J. Crowther, dr. O. Efthimiou, dr. T.L. Nguyen, dr. V.M.T. de Jong, dr. R.P. Venekamp, dr. T. Takada, M. van Barreveld, and M. Belias for the fruitful collaborations that I hope inspire future work.

I would like to thank the reading and opposing committee: prof. dr. S van Buuren, prof. dr. K.C.B. Roes, Dr. M. van Smeden, prof. dr. F.L.J. Visseren, dr. ir. D. van Klaveren, and prof. dr. A.M. May for taking the time to share their expertise and critically read and comment on this thesis.

Special thanks go out to prof. A.H. Zwinderman for helping me get on the right track with helpful advice on both statistical and career related topics. Likewise, I would like to thank dr. R.B. Geskus and prof. dr. ir. P.H.C. Eilers for their time and advise.

I wish to thank everyone at departments of Epidemiology & Health Economics and of Biostatistics at the Julius Center, and especially those involved in the Methods program, for being a very open, supportive, and active group of colleagues that was a great pleasure to work with. Toshi, thank you for the collaboration on numerous projects and for the time spent outside working hours.

I am grateful for prof. dr. J. Berkhof and prof. dr. ir. M.A. van de Wiel at the Epidemiology and Data Science department at the Amsterdam UMC for providing me with the opportunity to continue my research and for their support.

Tot slot dank ik mijn familie en vrienden. In het bijzonder dank ik mijn paranimfen Teun en Ruth, mijn zus Linda, mijn ouders, en mijn geliefde Iris voor alles door de jaren heen. Het is een geluk om jullie dichtbij te hebben. Janos en Jeanne, jullie vrolijkheid heeft me veel energie gegeven en jullie hebben me vaak thuisgebracht. Iris, je bent heel bijzonder en ik ben je bijzonder dankbaar voor veel meer dan hier te beschrijven is.