Beyond the average treatment effect

Risk-based approaches to medical decision making

Alexandros Rekkas

3/19/23

Table of contents

Preface

Introduction

Prediction of outcome risk

A patient's baseline risk—her or his probability of experiencing an outcome of interest—is a crucial component of medical decision making. For example, the patients' baseline 10-year cardiovascular is essential in the European Society of Cardiology and the European Society of Hypertension guidelines of 2018 for the management of arterial hypertension[1]. Similarly, an algorithm for management of osteoporosis has been suggested, based on a patient's osteoporotic fracture risk[2].

Risk prediction models are mathematical functions relating the presence of the outcome of interest to a set of measured predictors (covariates). These models are important tools for the assessment of a patient's baseline risk[3]. The performance of a prediction model in new patients is crucial. Model performance is often expressed by its discrimination, i.e. its ability to separate lower from higher risk patients, and its calibration, i.e. the agreement of predicted risk to observed event rates[4]. Although a risk prediction model may perform well in terms of discrimination and calibration for risk, it is not necessarily helpful for medical decision making. Baseline risk is one of the crucial pieces required for predicting individual responses to treatment. Knowledge of the patients' responsiveness to treatment, their vulnerability to side-effects and their preferences for other relevant outcomes is necessary information required for making truly informed clinical decisions[5]. Predicting more individualized responses to treatment is the main challenge of this thesis.

Prediction of treatment effect

In order to provide optimal medical care, doctors are advised to align their clinical practice with the results of well-conducted clinical trials, or the aggregated results from multiple such trials[6]. This approach implicitly assumes that all patients eligible for treatment experience the same effects—benefits and harms—of treatment as the reference trial population. However, the estimated treatment effect is often an average of heterogeneous treatment effects and, as such, may not be applicable to most patient subgroups, let alone individual patients. If a treatment causes a serious adverse event, then treating all patients on the basis of an observed average positive treatment effect may be harmful for some[7].

Heterogeneity of treatment effect is the variation of treatment effects on the individual level across the population[5]. The identification and quantification of heterogeneity of treatment effect is crucial for guiding medical decision making and lies at the core of patient-centered outcomes research. Despite heterogeneity of treatment effect being widely anticipated, its evaluation

is not straightforward. Individual treatment effects are—by their nature—unobservable: the moment a patient receives a specific treatment, their response under the alternatives becomes unmeasurable (the "fundamental problem of causal inference"[8]).

To evaluate a specific individual's response under an alternative treatment, researchers need to observe the outcomes of other "similar" patients that actually received the alternative treatment. More individualized treatment effects are often derived from the average effects estimated within a subgroup of similar patients. However, patient similarity is not straightforward to assess. Patients differ in a vast number of characteristics which may or may not be relevant to modifying treatment responses (the "reference class problem"[9]. Identification of such patient characteristics is challenging. In clinical trials it usually relies on the detection of statistically significant (pre-specified) interactions of treatment with measured covariates (subgroup analyses).

As clinical trials are in general only adequately powered to detect an overall treatment effect of a certain size, subgroup analyses can be highly problematic. Lack of statistical power often results in falsely concluding "consistency" of the treatment effect across several subpopulations of interest or overestimating the effect size of a treatment-covariate interaction. The former because an existing interaction effect was smaller than the detectable effect size, the latter because of false positives introduced from multiple testing. In Figure 1 the statistical power for detecting an interaction effect of equal size to the main effect is below 30%, despite the clinical trial being powered at 80% for the detection of the overall effect. Existing guidance on carrying out subgroup analyses attempts to mitigate these issues[10–12].

Prediction of treatment effect using outcome risk

Baseline risk is an important determinant of treatment effect[13,14]. It sets an upper bound on the treatment effect size. Low risk patients can only experience minimal treatment benefit before their risk is reduced to zero, while high risk patients can benefit much more (Figure 2). Consequently, baseline risk can be used as a subgrouping variable for assessing heterogeneity of treatment effect. For many populations of patients for whom we aim to estimate treatment effects, well-performing models for predicting baseline risk already exist and can be used to stratify the patients into subgroups (**REFS**). If no such models exist, the researcher can develop one in the dataset that is used for treatment effect estimation[13,15–18].

Baseline risk can also be used directly for predicting individual treatment

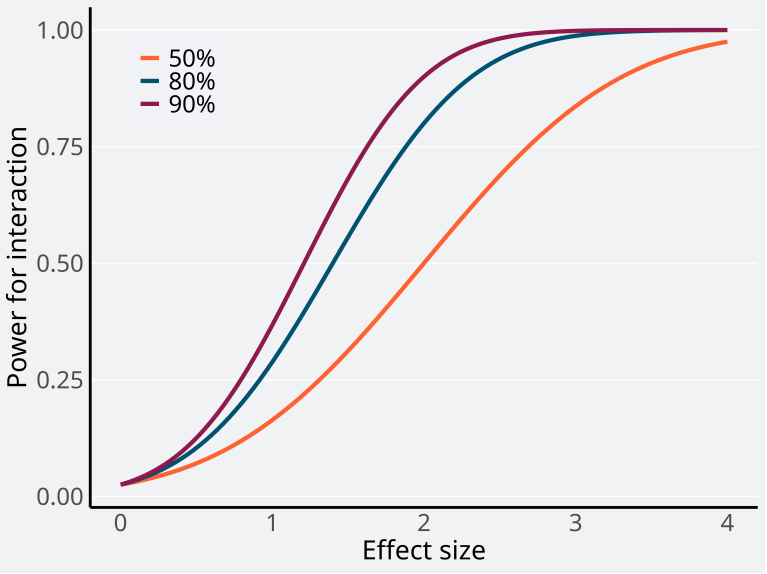

Figure 1: Statistical power for the detection of an interaction when the interaction effect size is between 0 and 4 times the main effect size. For simplicity, we assume equal number of treated and untreated patients and that patients are equally separated between the subgroup levels

benefit[19,20]. For example Califf et al[19] predicted individual benefits regarding 30-day mortality with tissue plasminogen activator (tPA) compared to streptokinase treatment in patients with acute myocardial infarction using baseline mortality risk and assuming a constant relative tPA treatment effect. However, relative treatment effect does not need to be constant. Modeling more flexible interactions of treatment with baseline outcome risk may provide more accurate absolute benefit predictions for individual patients.

Depending on the scale treatment effect is measured on, heterogeneity of treatment effect may or may not be identified (Figure 1.2). For example, despite finding statistically significant subgroup effect evaluated on the relative scale, the absolute risk difference between the two groups may be too small to have any clinical relevance[21]. Therefore, in the presence of a truly effective treatment, effect heterogeneity should always be anticipated on some scale[20], as baseline risk is bound to vary across trial patients. If effect modifiers are known and the available sample size provides adequate statistical power for evaluating treatment-covariate interactions, modeling these interactions would be the optimal approach for assessing heterogeneity of treatment effect. However, this approach may lead to overfitting and unstable estimates for the

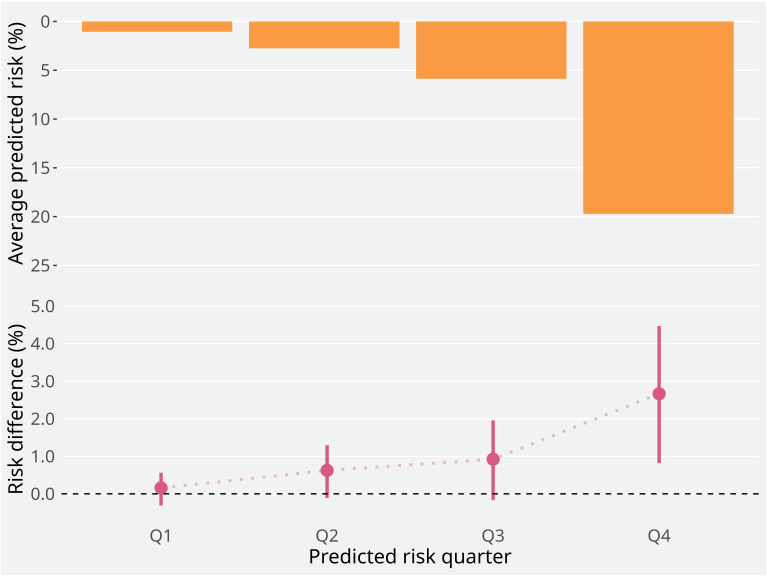

Figure 2: Statistical power for the detection of an interaction when the interaction effect size is between 0 and 4 times the main effect size. For simplicity, we assume equal number of treated and untreated patients and that patients are equally separated between the subgroup levels

interaction effects if the aforementioned conditions are not met[22].

Observational data

Healthcare data is routinely collected by general practitioners, hospitals, insurance companies, and many other private or public bodies and is becoming increasingly available, giving researchers access to massive amounts of patient data. Theoretically, the aforementioned statistical power challenges for the evaluation of heterogeneity of treatment effect would be largely mitigated if the analyses were performed on even a single such database. However, as this data is not being accumulated for research purposes, it suffers from many biases causing many commonly used methods to fail. Doctors prescribing a specific treatment expect—usually based on results from clinical trials—that it will be beneficial for the patient they are treating. This causes systematic differences in important characteristics among patients receiving different treatments and renders their comparison very challenging.

If all relevant patient characteristics on which the treating physician based their decision have been captured in the observational dataset, methods are available that can be used to account for these systematic differences[23–26].

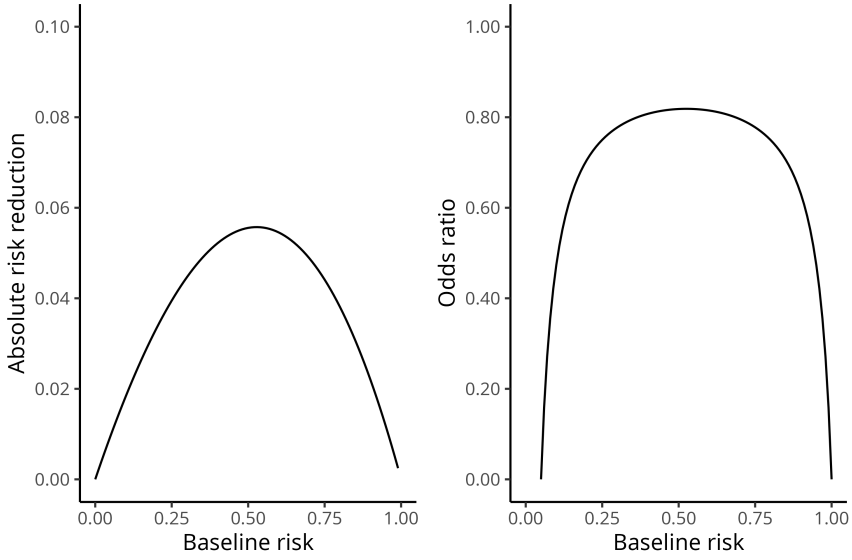

Figure 3: Scale dependency of treatment effect heterogeneity. In the left panel a constant odds ratio of 0.8 is assumed. In the right panel a constant absolute risk reduction of 0.1 is assumed.

Among the more popular ones is limiting the analyses to the propensity score matched subpopulation. Propensity scores are the patient-specific probabilities of receiving the treatment under study and have been shown to have the balancing property, that is, conditional on the propensity score, treatment assignment is independent of the potential outcomes[23]. This means that in a subset of patients with equal propensity scores, covariate distributions do not differ between patients receiving the treatment under study and those who are not. Consequently, patients within this subset can be assumed to be randomized.

Unfortunately, more often than not, a critical amount of the information that was used for treatment decisions is not captured. As a consequence, propensity score adjustment will not suffice to evaluate treatment effects using the observational data, be it overall or subgroup effects. Sensitivity analyses searching for evidence of this systematic unmeasured imbalances have been proposed and can be of assistance in many situations[27–30].

Another important problem with observational databases is that they use different architectures. As anyone gathering routinely observed healthcare data did so in a way that was more convenient to them, a plethora of structures for the resulting databases arose. Diseases, treatments, medical exams and many

more aspects of healthcare are often coded differently in different observational databases. In addition, more fundamental disparities between databases also factor in database incompatibility: different types of information are recorded in different databases. Different patient characteristics are captured—at different levels of detail—in a general practitioner database, in a hospital medical record or in an administrative claims database. This means that combining results from multiple databases is not a simple task

One of the solutions for handling database incompatibility was the creation of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)[31,32]. This provided a standard for structuring an observational database, while large effort was put into developing processes for mapping existing databases from their own specific structure to OMOP-CDM. With this high level of standardization achieved, the design and execution of highly scalable observational studies was made possible. Common definitions of diseases, treatments, and outcomes can now be applied uniformly across a network of many databases containing information on hundreds of millions of patients. An analysis plan can be executed following the exact same steps across the network providing effect estimates derived in different populations. The fragmented information scattered across multiple databases can now be summarized in a consistent way to give a fuller picture.

The power of the common database structure was demonstrated in a large-scale comparative effectiveness study of first-line treatment for hypertension[33]. This study compared five different first-line drug classes prescribed for hypertension regarding three primary effectiveness, six secondary effectiveness, and 46 safety outcomes across a global network of nine observational databases, all mapped to OMOP-CDM. The results complemented the already available evidence generated in clinical trials, confirming earlier findings and providing effect estimates on previously unexplored comparisons.

Observational databases provide access to millions of "real-life" patients. This motivates the exploration of methods for the assessment of treatment effect heterogeneity in the observational setting despite the challenges inherent to this type of data. The statistical power problem related to multiple subgroup analyses can still be present, as observational data is high-dimensional, i.e., the number of measured patient characteristics increases with the number of patients. Attempting a treatment effect modeling approach, where treatment-covariate interactions are modeled for the prediction of individualized treatment benefits, suffers from the same statistical power issues and often results in highly variable estimates. Therefore, using baseline outcome risk as the subgrouping variable, can provide useful insight into treatment effect heterogeneity within the observational setting. Modern libraries for developing risk predic-

tion models and for correcting for confounding are available and—capitalizing on OMOP-CDM—can be easily applied across databases with millions of patients.

Aims

The overall aim of this thesis is to explore the use of risk prediction models as the basis for medical decision making. We will study and apply methods for the evaluation of heterogeneity of treatment effect in both clinical trial data and observational data. The specific research aims are:

1. *Systematically review the current literature on predictive approaches to treatment effect heterogeneity.* The focus is on regression modeling approaches applied in clinical trial data.
2. *Develop scalable and reproducible risk-based predictive approaches to the assessment of heterogeneity of treatment effect.* We will explore new risk stratification approaches in observational settings and more individualized approaches in the clinical trial setting.
3. *Apply risk-based methods to better guide medical decisions.* We will develop baseline risk prediction models in several clinical case studies.

In Chapter 1 we present the results of a scoping literature review of regression modeling approaches for the assessment of treatment effect heterogeneity in the clinical trial setting. In Chapter 2 we develop a standardized scalable framework for the assessment of treatment effect heterogeneity using a risk-stratified approach in the observational setting. In Chapter 3 we compare different risk-based methods for predicting individualized treatment effects using extensive simulations of clinical trials. In Chapter 4 we develop and externally validate a model for the prediction of 5-year recurrence risk in sentinel node positive melanoma patients, using data from nine European Organization for Research and Treatment of Cancer centers. In Chapter 5 we develop and temporally validate a model for the prediction of 28-day mortality and admission to the ICU for patients presenting at the emergency department with suspected COVID-19 infection at four large Dutch hospitals between March and August, 2020. In Chapter 6 we apply the standardized framework developed in Chapter 2 to evaluate effect heterogeneity of teripatide treatment compared to oral bisphosphonates in female patients above the age of 50 with established osteoporosis. Finally, in Chapter 7 we present a general discussion along with perspectives on future work.

CHAPTER 1

Literature review

Chapter based on *Rekkas, A., Paulus, J.K., Raman, G. et al. Predictive approaches to heterogeneous treatment effects: a scoping review. BMC Med Res Methodol 20, 264 (2020). https://doi.org/10.1186/s12874-020-01145-1*

Abstract

Background: Recent evidence suggests that there is often substantial variation in the benefits and harms across a trial population. We aimed to identify regression modeling approaches that assess heterogeneity of treatment effect within a randomized clinical trial.

Methods: We performed a literature review using a broad search strategy, complemented by suggestions of a technical expert panel.

Results: The approaches are classified into 3 categories: 1) Risk-based methods (11 papers) use only prognostic factors to define patient subgroups, relying on the mathematical dependency of the absolute risk difference on baseline risk; 2) Treatment effect modeling methods (9 papers) use both prognostic factors and treatment effect modifiers to explore characteristics that interact with the effects of therapy on a relative scale. These methods couple data-driven subgroup identification with approaches to prevent overfitting, such as penalization or use of separate data sets for subgroup identification and effect estimation. 3) Optimal treatment regime methods (12 papers) focus primarily on treatment effect modifiers to classify the trial population into those who benefit from treatment and those who do not. Finally, we also identified papers which describe model evaluation methods (4 papers).

Conclusions: Three classes of approaches were identified to assess heterogeneity of treatment effect. Methodological research, including both simulations and empirical evaluations, is required to compare the available methods in different settings and to derive well-informed guidance for their application in RCT analysis.

Introduction

Evidence based medicine (EBM) has heavily influenced the standards of current medical practice. Randomized clinical trials (RCTs) and meta-analyses of RCTs are regarded as the gold standards for determining the comparative efficacy or effectiveness of two (or more) treatments within the EBM framework. Within this framework, as described in Guyatt et al's classic User's Guide to the Medical Literature II[10] "if the patient meets all the trial inclusion criteria, and doesn't violate any of the exclusion criteria—there is little question that the results of the trial are applicable". It has thus been argued that RCTs should attempt to include even broader populations to ensure generalizability of their results to more (and more diverse) individuals[34,35].

However, generalizability of an RCT result and applicability to a specific patient move in opposite directions[36,37]. When trial enrollees differ from one another in many observed determinants of the outcome of interest (both primary and safety), it can be unclear to whom the overall average benefit-harm trade-offs actually apply—even among those included in the trial[7,38]. Precision medicine aims to target the appropriate treatment to the appropriate patients. As such, analysis of heterogeneity of treatment effect, i.e. non-random variation in the direction or magnitude of a treatment effect for subgroups within a population[39], is the cornerstone of precision medicine; its goal is to predict the optimal treatments at the individual level, accounting for an individual's risk for harm and benefit outcomes.

In this scoping review[40], we aim to identify and categorize the variety of regression-based approaches for predictive heterogeneity of treatment effects analysis. Predictive approaches to analyses of heterogeneity of treatment effect are those that provide individualized predictions of potential outcomes in a particular patient with one intervention versus an alternative or, alternatively, that can predict which of 2 or more treatments will be better for a particular patient, taking into account multiple relevant patient characteristics. We distinguish these analyses from the typical onevariable-at-a-time subgroup analyses that appear in forest plots of most major trial reports, and from other analyses of heterogeneity of treatment effect which explore or confirm hypotheses regarding whether a specific covariate or biomarker modifies the effects of therapy. To guide future work on individualizing treatment decisions, we aimed to summarize the methodological literature on regression modeling approaches to predictive analysis of heterogeneity of treatment effect.

Methods

The terminology in this scoping review hews closely to that in the PATH Statement and PATH Statement Explanation and Elaboration articles, and we refer readers to these papers for details. Generally, we use the term heterogeneity of treatment effect to refer to a scale-dependent property. This is in distinction to other writers that have reserved the term heterogeneity of treatment effect to refer specifically to heterogeneity on a relative scale (***REF: 10***). Thus, when outcome risk varies across subgroups of patients, heterogeneity of treatment effect must exist on some scale. If relative risk is constant, then there is heterogeneity of treatment effect on the clinically important absolute scale. Nevertheless, since this review focuses on regression methods which are typically performed on the odds or hazard ratio scales, when we use the terms "effect modifier" and "effect modification" and "statistical interaction", we are generally referring to effect modification on a relative scale (e.g. hazard ratio or odds ratio), unless we otherwise specify—although we recognize that these too are scale dependent concepts[14,16,17,20,41]. Additionally, we note that we generally eschew the term "individual treatment effects", since person level effects cannot be observed or measured in parallel arm clinical trials (owing to the fundamental problem of causal inference, only one counterfactual outcome can be observed in a given patient). Nevertheless, the common goal of the different methods of predictive approaches to heterogeneity of treatment effect we describe herein is to provide "*individualized*" treatment effect estimates from groupbased data, since medical decisions are generally made at the individual person level[20]. These treatment effects are estimated conditional on many covariates, which are felt to be relevant for determining the benefits of therapy.

Due to the absence of medical subject headings (MeSH) for heterogeneity of treatment effect, we used a relatively broad search strategy to maximize sensitivity. For the time period 1/1/2000 through 8/9/2018, we searched Medline and Cochrane Central using the text word search strategy from Table 1.1. We also retrieved seminal articles suggested by a technical expert panel (TEP). The TEP was comprised of 16 experts who represented various perspectives on predictive analyses of heterogeneity of treatment effect, including treatment effect heterogeneity, prediction modeling, clinical trials, and guideline development as well as a patient advocate. More details on the TEP are available in the PATH Statement[16,17].

We sought papers that developed or evaluated methods for predictive heterogeneity of treatment effect in the setting of parallel arm RCT designs or simulated RCT. Abstracts were screened to identify papers that developed or

#	Results
1	((heterogen$ and effect$) or (effect and modif$)).tw.
2	regression.tw.
3	treatment$.tw.
4	(treatment adj1 effect$).tw.
5	(treatment adj1 difference$).tw.
6	exp risk/ or risk.tw.
7	3 or 4 or 5 or 6
8	*Models, Statistical/
9	*Randomized Controlled Trials as Topic/mt
10	Multicenter Studies as Topic/mt
11	*Randomized Controlled Trials as Topic/sn
12	Multicenter Studies as Topic/sn
13	*Clinical Trials as Topic/sn
14	*Precision Medicine/mt
15	or/8–14
16	1 and 2
17	2 and 7
18	15 and 17
19	15 and 16
20	18 or 19

Table 1.1: Search strategy for the study.

evaluated a regression-based method for predictive treatment effect heterogeneity on actual or simulated parallel arm RCT data. Papers describing a generic approach that could be applied using either regression or non-regression methods, or papers comparing regression to non-regression methods were also included. Similarly, papers comparing generic one-variable-at-a-time approaches to predictive heterogeneity of treatment effect methods were also included. Finally, papers suggested by the TEP that fell outside the search window were considered for inclusion.

We excluded papers solely related to cross-over, single-arm, and observational study designs. We also excluded papers that were primarily applications of existing methods, such as those that primarily aim to estimate a treatment effect of interest in a specific patient population, rather than papers with the primary aim of developing or evaluating methods of predictive heterogeneity of treatment effect analysis. We also excluded papers using only non-regression-based methods. Similarly, methods papers about ONLY non-predictive subgroup

analysis, i.e. one-variable-at-a-time or conventional subgroups, were omitted. We excluded papers on trial enrichment or adaptive trial designs along with those that use predictive heterogeneity of treatment effect approaches in the design. We also excluded papers primarily aiming at characterization or identification of heterogeneity in response rather than trying to predict responses for individual patients or subsets of patients; e.g. group based trajectory or growth mixture modeling. Papers on regression methods that make use of covariates post-baseline, or temporally downstream of the treatment decision were omitted. Review articles and primarily conceptual papers without accompanying methods development were also excluded.

Titles, abstracts and full texts were retrieved and double-screened by six independent reviewers against eligibility criteria. Disagreements were resolved by group consensus in consultation with a seventh senior expert reviewer (DMK) in meetings.

Results

We identified 2510 abstracts that were screened in duplicate. We retrieved 64 full-text articles and an additional 110 suggested by experts and identified from reference lists of eligible articles. These 174 full-text articles were again screened in duplicate with group consensus resolution of conflicts in meetings. A total of 36 articles met eligibility criteria (Figure 1.1).

Categorization methods

We could classify all regression-based methods to predictive heterogeneity of treatment effect into 3 broad categories based on whether and how they incorporated prognostic variables and relative treatment effect modifiers:

- Risk-based methods exploit the mathematical dependency of treatment benefit on a patient's baseline risk for the outcome under study[39,40]. Even though relative treatment effect may vary across different levels of baseline risk, relative treatment effect modification by each covariate is not considered, i.e. no covariate by treatment interaction terms are considered (***Table 2, eqs. 1–3***).
- Treatment effect modeling methods use both the main effects of risk factors and covariate-by -treatment interaction terms (on the relative scale) to estimate individualized benefits. They can be used either for making individualized absolute benefit predictions or for defining patient subgroups with similar expected treatment benefits (***Table 2, eq. 4***).
- Optimal treatment regime methods focus primarily on treatment effect modifiers (covariate by treatment interactions) for the definition of a treatment assignment rule dividing the trial population into those who

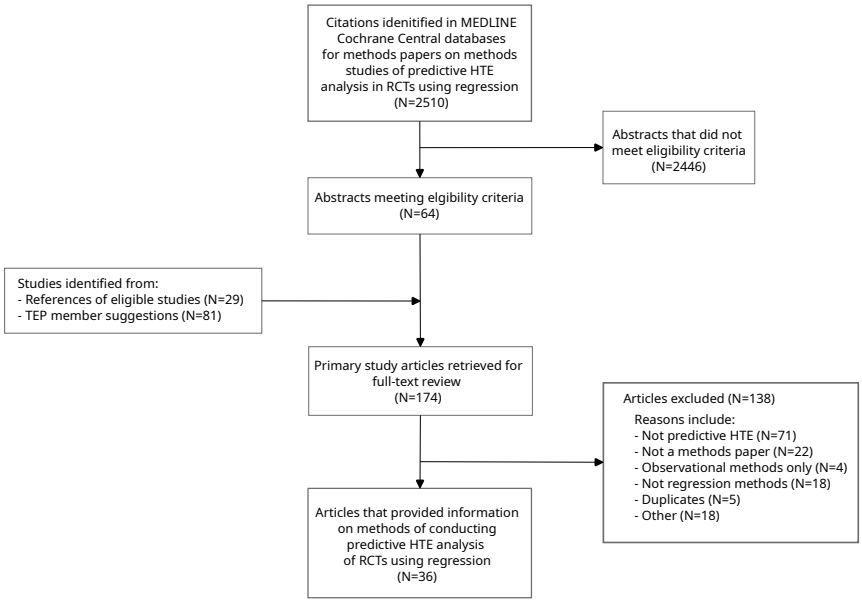

Figure 1.1: Figure flow chart

benefit from treatment and those who do not (***Table 2, eq. 5***). Contrary to previous methods, baseline risk or the magnitude of absolute treatment benefit are not of primary concern.

Although risk-based methods emerged earlier (***Fig. 2***), methodology papers on treatment effect modeling (9 papers) and optimal treatment regimes (12 papers) are more frequently published since 2010 than papers on risk-based methods (8 papers). Even though extensive literature exists on model evaluation when it comes to prediction modeling, the same task can be quite challenging when modeling treatment effects[4]. That is due to the unavailability of counterfactual outcomes under the alternative treatment, providing a substantial challenge to the assessment of model fit. Methods included in the review concerning model evaluation in the setting of predictive heterogeneity of treatment effect (4 papers) were assigned to a separate category as they are relevant to all identified approaches.

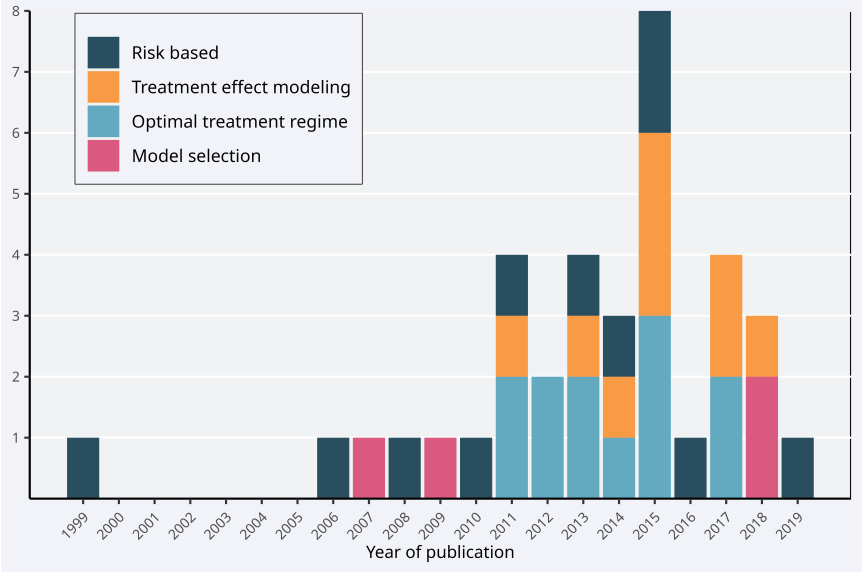

Figure 1.2: Results of the review.

Box 1.1: Equations corresponding to treatment effect hetergoenetiy assessment methods.

Risk modeling

A multivariate regression model f that predicts the risk of an outcome y based on the predictors $x_1, \dots, x_p$ is identified or developed:

$$\text{risk}(x_1, \dots, x_p) = E\{y \mid x_1, \dots, x_p\} = f(\alpha + \beta_1 x_1 + \dots \beta_p x_p)$$

The expected outcome of a patient with measured predictors $x_1, \dots x_p$ receiving treatment T (where $T = 1$, when patient is treated and 0 otherwise) based on the linear predictor $\text{lp}(x_1, \dots, x_p) = \alpha + \beta_1 x_1 + \dots + \beta_p x_p$ from a previously derived risk model can be described as:

$$E\{y \mid x_1, \dots, x_p\} = f(lp + \gamma_0 T + \gamma_1 T \times lp) \tag{1.1}$$

When the assumption of constant relative treatment effect across the entire risk distribution is made (risk magnification) equation 1.1 takes the form:

$$E\{y \mid x_1, \dots, x_p\} = f(lp + \gamma_0 T)$$

Treatment effect modeling

The expected outcome of a patient with measured predictors $x_1, \ldots, x_p$ receiving treatment T can be derived from a model containing predictor main effects and potenetial treatment interaction terms:

$$E\{y \mid x_1, \ldots, x_p\} = f(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \gamma_0 T + \gamma_1 T x_1 + \cdots + \gamma_p T x_p)$$

Optimal treatment regime
A treatment regime $T(x_1, \ldots, x_p)$ is a binary treatment assignment rule based on measured predictors. The optimal treatment regime maximizes the overall expected outcome across the entire target population:

$$T_{\text{optimal}} = \text{argmax}_T E\Big\{E\{y \mid x_1, \ldots, x_p, T(x_1, \ldots, x_p)\}\Big\}$$

Risk-based methods

The most rigid and straightforward risk-based methods assume a constant relative treatment effect across different levels of baseline risk and ignore potential interactions with treatment. Dorresteijn et al.[42] studied individualized treatment with rosuvastatin for prevention of cardiovascular events. They combined existing prediction models (Framingham score, Reynolds risk score) with the average rosuvastatin effect found in an RCT. To obtain individualized absolute treatment benefits, they multiplied baseline risk predictions with the average risk reduction found in trials. The value of the proposed approach is assessed in terms of improved decision making by comparing the net benefit with treat-none and treat-all strategies[43]. Julien and Hanley[44] estimated prognostic effects and treatment effect directly form trial data, by incorporating a constant relative treatment effect term in a Cox regression model. Patient-specific benefit predictions followed from the difference between event-free survival predictions for patients with and without treatment. A similar approach was used to obtain the predicted 30- day survival benefit of treatment with aggressive thrombolysis after acute myocardial infarction[19].

Risk stratification approaches analyze relative treatment effects and absolute treatment effects within strata of predicted risk, rather than assuming a constant relative effect. Both Hayward et al.[45] and Iwashyna et al.[46] demonstrated that these methods are useful in the presence of treatment-related harms to identify patients who do not benefit (or receive net harm) from a treatment that is beneficial on average. In a range of plausible scenarios evaluating heterogeneity of treatment effect when considering binary endpoints, simulations showed that studies were generally underpowered to detect covariate-by-

treatment interactions, but adequately powered to detect risk-by-treatment interactions, even when a moderately performing prediction model was used to stratify patients. Hence, risk stratification methods can detect patient subgroups that have net harm even when conventional methods conclude consistency of effects across all major subgroups.

Kent et al.[13] proposed a framework for the analysis of heterogeneity of treatment effect in RCT data that recommended published trials routinely report the distribution of baseline risk in the overall study population and in the separate treatment arms using a risk prediction tool. Primarily binary or time-to-event outcomes were considered. Researchers should demonstrate how relative and absolute risk reduction vary by baseline risk and test for heterogeneity of treatment effect with interaction tests. Externally validated prediction models should be used, when available.

In the absence of an adequate prediction model when performing a risk-based assessment of heterogeneity of treatment effect, an internal risk model from the data at hand can be derived. Burke et al.[15] demonstrated that developing the risk model on the control arm of the trial may result in overfitting and, thus, exaggerate the presence of heterogeneity of treatment effect. In extensive simulations, internally developed prediction models blinded to treatment assignment led to unbiased treatment effect estimates in strata of predicted risk. Using this approach to re-analyze 32 large RCT, Kent et al.[47] demonstrated that variation in the outcome risk within an RCT is very common, in the presence of adequately performing prediction models, which in turn leads to substantial heterogeneity of treatment effect on the clinically important scale of absolute risk difference. Several trials from this analysis had clinically relevant results[48–50].

Similar to Burke et al.[15], Abadie et al.[18] presented evidence of large biases in risk stratified assessment of heterogeneity of treatment effect in two randomized experiments rising from the development of a prediction model solely from the control arm. They focused on financial outcomes that are primarily continuous. As a remedy, they considered both a leave-one-out approach, where individualized risk predictions are obtained from a model derived by excluding the particular individual, and a repeated split sample approach, where the original sample is repeatedly split into a sample for the development of the prediction model and a sample for treatment effect estimation within risk strata. These approaches were found to substantially reduce bias in a simulation study. Finally, Groenwold et al.[51] found in simulations that the inclusion of a constant relative treatment effect in the development of a prediction model better calibrates predictions to the untreated population. However, this approach may not be optimal for risk-based assessment of treatment effect heterogene-

ity, where accurate ranking of risk predictions is of primary importance for the calibration of treatment benefit predictions.

Follmann and Proschan[52] proposed a one-step likelihood ratio test procedure based on a proportional interactions model to decide whether treatment interacts with a linear combination of baseline covariates. Their proportional interactions model assumes that the effects of prognostic factors in the treatment arm are equal to their effects in the control arm multiplied by a constant, the proportionality factor. Testing for an interaction along the linear predictor amounts to testing that the proportionality factor is equal to 1. If high risk patients benefit more from treatment (on the relative scale) and disease severity is determined by a variety of prognostic factors, the proposed test results in greater power to detect heterogeneity of treatment effect on the relative scale compared to multiplicity-corrected subgroup analyses. Even though the proposed test requires a continuous response, it can be readily implemented in large clinical trials with binary or time-toevent endpoints.

Kovalchik et al.[53] expanded upon the previous approach by exploring misspecification of the proportional interactions model, when considering a fixed set of pre-specified candidate effect modifiers. A proportional interactions model is miss-specified either when covariates with truly proportional effects are excluded or when covariates with non-proportional effects across treatment arms are included in the model. In this case the one-step likelihood ratio test of Follmann and Proschan[52] fails to achieve its statistical advantages. For model selection an all subsets approach combined with a modified Bonferroni correction method can be used. This approach accounts for correlation among nested subsets of considered proportional interactions models, thus allowing the assessment of all possible proportional interactions models while controlling for the familywise error rate.

Treatment effect modeling

Using data from the SYNTAX trial[54] Van Klaveren et al.[55] considered models of increasing complexity for the prediction of heterogeneous treatment effects using data from the SYNTAX trial. They compared different Cox regression models for the prediction of treatment benefit: 1) a model without any risk factors; 2) a model with risk factors and a constant relative treatment effect; 3) a model with treatment, a prognostic index and their interaction; and 4) a model including treatment interactions with all available prognostic factors, fitted both with conventional and with penalized ridge regression. Benefit predictions at the individual level were highly dependent on the modeling strategy, with treatment interactions improving treatment recommendations under certain circumstances.

Basu et al.[56] developed and validated risk models for predicting the absolute benefit (reduction of time to CVD events) and harm (serious adverse events) from intensive blood pressure therapy, using data from SPRINT. They compared traditional backward selection to an elastic net approach for selection and estimation of all treatment-covariate interactions. The two approaches selected different treatment-covariate interactions and while their performance in terms of CVD risk prediction was comparable when externally validated in the ACCORD BP trial[57]—the traditional approach performed considerably worse than the penalized approach when predicting absolute treatment benefit. However, with regard to selection of treatment interactions, Ternes et al.[58] concluded from an extensive simulation study that no single methodology yielded uniformly superior performance. They compared 12 different approaches in a high-dimensional setting with survival outcomes. Their methods ranged from a straightforward univariate approach as a baseline, where Wald tests accounting for multiple testing were performed for each treatment-covariate interaction to different approaches for dealing with hierarchy of effects—whether they enforce the inclusion of the respective main effects if an interaction is selected—and also different magnitude of penalization of main and interaction effects.

Another approach to reducing overfitting of treatment effect models is separation of treatment effect estimation from subgroup identification. Cai et al.[59] fit "working" regression parametric models within treatment arms to derive absolute treatment benefit scores initially. In a second stage, the population is stratified into small groups with similar predicted benefits based on the firststage scores. A non-parametric local likelihood approach is used to provide a smooth estimate of absolute treatment benefit across the range of the derived sores. The authors focused on continuous and binary endpoints, but their method can be extended to time-to-event outcomes. Claggett et al.[60] extended this two-stage methodology to RCTs with multiple outcomes, by assigning outcomes into meaningful ordinal categories. Overfitting can be avoided by randomly splitting the sample into two parts; the first part is used to select and fit ordinal regression models in both the treatment and the control arm. In the second part, the models that perform best in terms of a cross-validated estimate of concordance between predicted and unobservable true treatment difference— defined as the difference in probability of observing a worse outcome under control compared to treatment and the probability of observing a worse outcome under treatment compared to control—are used to define treatment benefit scores for patients. Treatment effects conditional on the treatment benefit score are then estimated through a nonparametric kernel estimation procedure.

Zhao et al.[61] proposed a two-stage methodology similar to the approach of Cai

et al.[59], focusing on the identification of a subgroup that benefits from treatment. They repeatedly split the sample population based on the first-stage treatment benefit scores and estimate the treatment effect in subgroups above different thresholds. These estimates are plotted against the score thresholds to assess the adequacy of the selected scoring rule. This method could also be used for the evaluation of different modeling strategies by selecting the one that identifies the largest subgroup with an effect estimate above a desired threshold.

Künzel et al.[62] proposed an "X-learner" for settings where one treatment arm is substantially larger than the alternative. They also start by fitting separate outcome models within treatment arms. However, rather than using these models to calculate treatment benefit scores, they imputed individualized absolute treatment effects, defined as the difference between the observed outcomes and the expected counterfactual (potential) outcomes based on model predictions. In a second stage, two separate regression models—one in each treatment arm— are fitted to the imputed treatment effects. Finally, they combined these two regression models for a particular covariate pattern by taking a weighted average of the expected treatment effects.

Most effect modeling methods start with outcome predictions conditional on treatment and then examine the difference in predictions with and without treatment. In contrast, Weisberg and Pontes[63] introduced a causal difference outcome variable ("cadit") which can be modeled directly. In case of a binary outcome, the binary cadit is 1 when a treated patient has a good outcome or when an untreated patient does not, and 0 otherwise. Thus, the dependent variable implicitly codes treatment assignment and outcome simultaneously. They first demonstrated that the absolute treatment benefit equals $2 \times \mathrm{P}(\text{cadit} = 1) - 1$ and then they derived patient-specific treatment effect estimates by fitting a logistic regression model to the cadit. A similar approach was described for continuous outcomes with the continuous cadit defined as -2 and 2 times the centered outcome, i.e. the outcome minus the overall average outcome, for untreated and treated patients, respectively.

Finally, Berger et al.[64] proposed a Bayesian methodology for the detection of subgroup treatment effects in case of a continuous response and binary covariates. The approach identifies single covariates likely to modify treatment effect, along with the expected individualized treatment effect. The authors also extended their methodology to include two covariates simultaneously, allowing for the assessment of multivariate subgroups.

Optimal treatment regimes

A treatment regime (TR) is a function mapping each patient's covariate pattern to a single treatment assignment. Any candidate TR can be evaluated based on its value, i.e. the expected outcome at the population level if the specific TR were to be followed. The TR achieving the highest value among all possible TRs is the optimal treatment regime (OTR). The majority of such methods follows a two-stage approach, where an outcome model—usually including treatment interactions—is used to derive expected treatment benefit in the first stage. In the second stage treatment assignment is optimized based on the expected outcome. Qian and Murphy[65] advocated a first-stage model including all covariate main effects and treatment interactions in combination with LASSO-penalization to reduce model complexity. Real-valued (continuous or binary) are considered without considering censoring.

When the outcome model is misspecified, however, the approach of Qian and Murphy may fail to identify the best possible treatment regime. Zhang et al[66] introduced an approach robust to such misspecifications that uses an augmented inverse probability weighted estimator of the value function. This is achieved by imposing a missing data framework, where the response under any candidate OTR is observed if the proposed treatment coincides with actual treatment and is considered missing otherwise. However, in commenting on this work, Taylor et al.[67] noted that the misspecification issues of the outcome models considered in the simulation study presented by Zhang et al. would have been easily spotted, if common approaches for the assessment of model fit had been examined. They argue that if adequately fitting outcome models had been thoroughly sought, the extra modeling required for the robust methods of Zhang et al. may not have been necessary.

Zhang et al.[68] proposed a novel framework for the derivation of OTRs for real-valued responses (continuous or binary), within which treatment assignment is viewed as a classification problem. The OTR is derived in two separate steps. In the first step, a contrast function is estimated, determining the difference between expected outcomes under different treatment assignments for each individual patient. The sign of the contrast function is then used to define class labels, i.e. -1 for negative contrast (harm) and + 1 for positive contrast (benefit). In the second step, any classification technique can be used to find the OTR by minimizing the expected miss-classification error weighted by the absolute contrast. The authors demonstrated that many of the already existing OTR methods[65,68] fit within their framework by defining a specific contrast function.

When the outcome of interest is continuous, the magnitude of absolute treat-

ment benefit estimates derived from regression-based methods depends solely on treatment interactions. Therefore, Foster et al.[69] focus on non-parametric estimation of the function defining the structure of treatment-covariate interactions for a continuous outcome of interest. More specifically, they recursively update non-parametric estimates of the treatment-covariate interaction function from baseline risk estimates and vice-versa until convergence. The estimates of absolute treatment benefit are then used to restrict treatment to a contiguous sub-region of the covariate space.

Xu et al.[70] claimed that the identification of an OTR with high value depends on the adequate assessment of the sign of treatment-covariate interactions rather than on the estimation of the contrast function. They demonstrated that in many common cases (binary or time-to-event outcomes), even though the underlying structure of interactions can be quite complex, its sign can be approximated from a much simpler linear function of effect modifiers. Using the jlassification framework of Zhang et al.[68], they assign patients to class labels based on the resulting sign from these candidate linear combinations. The coefficients of that linear function are derived by minimizing the misclassification error weighted by the observed outcome—assuming higher values are preferable. In this way, the derived OTR is forced to contradict actual treatment assignment when the observed outcome is low. Tian et al.[71] proposed a different approach that solely focuses on treatment-covariate interactions by recoding the binary treatment indicator variable to $-\frac{1}{2}$ for control patients and $+\frac{1}{2}$ for treated patients and multiplying it with the covariates of a posited regression model to derive modified covariates so that the linear predictor of the model predicting the outcome from the modified covariates can be used as a score for stratifying patients with regard to treatment benefits. Starting from continuous responses they generalized their methodology to binary and time-to-event outcomes.

Kraemer[72] suggested a methodology that implicitly assesses treatment-covariate interactions using the correlation coefficient of the pairwise difference of the continuous outcome between treatment arms and their respective candidate predictive factor pairwise difference as a measure of effect modification. A stronger composite treatment effect modifier can then be constructed by fitting a regression model predicting pairwise outcome differences between treatments from the averages of the effect modifier values across treatment arms and then summing the individual effect modifiers weighted by the estimated regression coefficients. Treatment can then be assigned based on stratification on the composite treatment effect moderator. Two different approaches to model selection in Kraemer's effect modifier combination method were identified in clinical applications. Principal component

analysis was used to select an uncorrelated subset from a large set of possibly correlated effect modifiers[73]. Alternatively, the cross-validated mean squared error of increasingly complex regression models was used to select the number of effect modifiers to construct the composite one[74].

Gunter et al.[75] proposed a method for the discovery of covariates that qualitatively interact with treatment. Using LASSO regression to reduce the space of all possible combinations of covariates and their interaction with treatment to a limited number of covariate subsets, their approach selects the optimal subset of candidate covariates by assessing the increase in the expected response from assigning based on the considered treatment effect model, versus the expected response of treating everyone with the treatment found best from the overall RCT result. The considered criterion also penalizes models for their size, providing a tradeoff between model complexity and the increase in expected response. The method focuses solely on continuous outcomes, however, suggestions are made on its extension to binary type of outcomes.

Finally, Petkova et al.[76] proposed to combine baseline covariates into a single generated effect modifier (GEM) based on the linear model. The GEM is defined as the linear combination of candidate effect modifiers and the objective is to derive their individual weights. This is done by fitting linear regression models within treatment arms where the independent variable is a weighted sum of the baseline covariates, while keeping the weights constant across treatment arms. The intercepts and slopes of these models along with the individual covariate GEM contributions are derived by maximizing the interaction effect in the GEM model, or by providing the best fit to the data, or by maximizing the statistical significance of an F-test for the interaction effects—a combination of the previous two. The authors derived estimates that can be calculated analytically, which makes the method easy to implement.

A growing literature exists on estimating the effect of introducing the OTR to the entire population[77–80]. Luedtke and Van der Laan[77] provide an estimate of the optimal value—the value of the OTR—that is valid even when a subset of covariates exists for which treatment is neither beneficial nor harmful. It has been previously demonstrated that estimation of the optimal value is quite difficult in those situations[81]. Based on the proposed method, an upper bound of what can be hoped for when a treatment rule is introduced can be established. In addition, Luedtke and Van der Laan[80] provided an estimation method for the impact of treating the optimal subgroup, i.e. the subgroup that is assigned treatment based on the OTR. Their methodology returns an estimate of the population level effect of treating based on the OTR compared to treating no one.

Model evaluation

Schuler et al.[82] defined three broad classes of metrics relevant to model selection when it comes to treatment effect modeling. -risk metrics evaluate the ability of models to predict the outcome of interest conditional on treatment assignment. Treatment effect is either explicitly modeled by treatment interactions or implicitly by developing separate models for each treatment arm. -risk metrics focus directly on absolute treatment benefit. However, since absolute treatment benefit is unobservable, it needs to be estimated first. Value-metrics originate from OTR methods and evaluate the outcome in patients that were assigned to treatment in concordance with model recommendations.

Vickers et al.[43] suggested a methodology for the evaluation of models predicting individualized treatment effects. The method relies on the expression of disease-related harms and treatment-related harms on the same scale. The minimum absolute benefit required for a patient to opt for treatment (treatment threshold) can be viewed as the ratio of treatment-related harms and harms from disease-related events, thus providing the required relationship. Net benefit is then calculated as the difference between the decrease in the proportion of disease-related events and the proportion of treated patients multiplied by the treatment threshold. The latter quantity can be viewed as harms from treatment translated to the scale of disease-related harms. Then, the net benefit of a considered prediction model at a specific treatment threshold can be derived from a patient-subset where treatment received is congruent with treatment assigned based on predicted absolute benefits and the treatment threshold. The model's clinical relevance is derived by comparing its net benefit to the one of a treat-all policy.

Van Klaveren et al.[22] defined a measure of discrimination for treatment effect modeling. A model's ability to discriminate between patients with higher or lower benefits is challenging, since treatment benefits are unobservable in the individual patient (since only one of two counterfactual potential outcomes can be observed). Under the assumption of uncorrelated counterfactual outcomes, conditional on model covariates, the authors matched patients from different treatment arms by their predicted treatment benefit. The difference of the observed outcomes between the matched patient pairs (0, 1: benefit; 0, 0 or 1, 1: no effect; 1, 0: harm) acts as a proxy for the unobservable absolute treatment difference. The c-statistic for benefit can then be defined on the basis of this tertiary outcome as the proportion of all possible pairs of patient pairs in which the patient pair observed to have greater treatment benefit was predicted to do so.

Finally, Chen et al.[83] focused on the case when more than one outcomes—

often non-continuous— are of interest and proposed a Bayesian model selection approach. Using a latent variable methodology, they link observed outcomes to unobservable quantities, allowing for their correlated nature. To perform model selection, they derive posterior probability estimates of false inclusion or false exclusion in the final model for the considered covariates. Following the definition of an outcome-space sub-region that is considered beneficial, individualized posterior probabilities of belonging to that beneficial sub-region can be derived as a by-product of the proposed methodology.

Discussion

We identified 36 methodological papers in recent literature that describe predictive regression approaches to the analysis of heterogeneity of treatment effect in RCT data. These methodological papers aimed to develop models for predicting individual treatment benefit and could be categorized as follows: 1) risk modeling (n = 11), in which RCT patients were stratified or grouped solely on the basis of prognostic models; 2) effect modeling (n = 9), in which patients are grouped or stratified by models combining prognostic factors with factors that modify treatment effects on the relative scale (effect modifiers); 3) optimal treatment regimes (n = 12), which seek to classify patients into those who benefit and those who do not, primarily on the basis of effect modifiers. Papers on the evaluation of different predictive approaches to heterogeneity of treatment effect (n = 4) were assigned to a separate category. Of note, we also found literature on the evaluation of biomarkers for treatment selection, which did not meet inclusion criteria[84–87].

Risk-based approaches use baseline risk determined by a multivariate equation to define the reference class of a patient as the basis for predicting heterogeneity of treatment effect. Two distinct approaches were identified: 1) risk magnification assumes constant relative treatment effect across all patient subgroups, while 2) risk stratification analyzes treatment effects within strata of predicted risk. This approach is straightforward to implement, and may provide adequate assessment of treatment effect heterogeneity in the absence of strong prior evidence for potential effect modification. The approach might better be labeled 'benefit magnification', since benefit increases by higher baseline risk and a constant relative risk.

Treatment effect modeling methods focus on predicting the absolute benefit of treatment through the inclusion of treatment-covariate interactions alongside the main effects of risk factors. However, modeling such interactions can result in serious overfitting of treatment benefit, especially in the absence of well-established treatment effect modifiers. Penalization methods such as LASSO regression, ridge regression or a combination (elastic net penaliza-

tion) can be used as a remedy when predicting treatment benefits in other populations. Staging approaches starting from—possibly overfitted— "working" models predicting absolute treatment benefits that can later be used to calibrate predictions in groups of similar treatment benefit provide another alternative. While these approaches should yield well calibrated personalized effect estimates when data are abundant, it is yet unclear how broadly applicable these methods are in conventionally sized randomized RCTs. Similarly, the additional discrimination of benefit of these approaches compared to the less flexible risk modeling approaches remains uncertain. Simulations and empirical studies should be informative regarding these questions.

The similarity of OTRs to general classification problems—finding an optimal dichotomization of the covariates space—enables the implementation of several existing non-regression-based classification algorithms. For instance Zhao et al.[88] applied a support vector machine methodology for the derivation of an OTR for a binary outcome and was later extended to survival outcomes[89]. Because prognostic factors do not affect the sign of the treatment effect, several OTR methods rely primarily on treatment effect modifiers. However, when treatments are associated with adverse events or treatment burdens (such as costs) that are not captured in the primary outcome—as is often the case—estimates of the magnitude of treatment effect are required to ensure that only patients above a certain expected net benefit threshold (i.e. outweighing the harms and burdens of therapy) are treated. Similarly, these classification methods do not provide comparable opportunity for incorporation of patient values and preferences for shared decision making which prediction methods do.

While there is an abundance of proposed methodological approaches, examples of clinical application of prediction models for treatment effect heterogeneity remain quite rare. This may reflect the fact that all these approaches confront the same fundamental challenges. These challenges include the unobservability of individual treatment response, the curse of dimensionality from the large number of covariates, the lack of prior knowledge about the causal molecular mechanisms underlying variation in treatment effects and the relationship of these mechanism to observable variables, and the very low power in which to explore interactions. Because of these challenges there might be very serious constraints on the usefulness of these methods as a class; while some methods may be shown to have theoretical advantages, the practical import of these theoretical advantages may not be ascertainable.

The methods we identified here generally approach the aforementioned challenges from opposite ends. Relatively rigid methods, such as risk magnification (in which relative effect homogeneity is assumed) and risk modeling (which

examines changes in relative effect according to baselines risk only) deal with dimensionality, low power and low prior knowledge by restricting the flexibility of the models that can be built to emphasize the well understood influence of prognosis. Effect modeling approaches permit more flexible modeling and then subsequently try to correct for the overfitting that inevitably arises. Based on theoretical considerations and some simulations, it is likely that the optimal approach depends on the underlying causal structure of the data, which is typically unknown. It is also likely that the method used to assess performance may affect which approach is considered optimal. For example, recent simulations have favored very simple approaches when calibration is prioritized, but more complex approaches when discrimination is prioritized—particularly in the presence of true effect modification[90]. Finally, it is uncertain whether any of these approaches will add value to the more conventional EBM approach of using an overall estimate of the main effect, or to the risk magnification approach of applying that relative estimate to a risk model.

We identify several limitations to our study. Because no MeSH identifying these methods exists, we anticipate that our search approach likely missed some studies. In addition, a recently growing literature of other non-regression based methods that assess predictive approaches to the assessment of treatment effect heterogeneity in observational databases[91–93] would have been excluded. Finally, our review is descriptive and did not compare the approaches for their ability to predict individualized treatment effects or to identify patient subgroups with similar expected treatment benefits.

Based on the findings and the limitations of our review, several objectives for future research can be described. Optimal approaches to the reduction of overfitting through penalization need to be determined, along with optimal measures to evaluate models intended to predict treatment effect. General principles to judge the adequacy of sample sizes for predictive analytic approaches to heterogeneity of treatment effect are required to complement the previous objectives. Also, methods that simultaneously predict multiple risk dimensions regarding both primary outcome risks and treatment-related harms need to be explored. The current regression-based collection of methods could be expanded by a review of non-regression approaches. Methods targeted at the observational setting need also to be considered. Additionally, a set of empirical and simulation studies should be performed to evaluate and compare the identified methods under settings representative of real world trials. The growing availability of publicly available randomized clinical trials should support this methodological research[94–96].

In conclusion, we identified a large number of methodological approaches for the assessment of heterogeneity of treatment effects in RCTs developed in the

past 20 years which we managed to divide into 3 broad categories. Extensive simulations along with empirical evaluations are required to assess those methods' relative performance under different settings and to derive wellinformed guidance for their implementation. This may allow these novel methods to inform clinical practice and provide decision makers with reliable individualized information on the benefits and harms of treatments. While we documented an exuberance of new methods, we do note a marked dearth of comparative studies in the literature. Future research could shed light on advantages and drawbacks of methods in terms of predictive performance in different settings.

CHAPTER 2

A framework for risk-based assessment of treatment effect heterogeneity

Chapter based on *Rekkas, A., van Klaveren, D., Ryan, P.B. et al. A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases. npj Digit. Med. 6, 58 (2023). https://doi.org/10.1038/s41746-023-00794-y*

Abstract

Treatment effects are often anticipated to vary across groups of patients with different baseline risk. The Predictive Approaches to Treatment Effect Heterogeneity (PATH) statement focused on baseline risk as a robust predictor of treatment effect and provided guidance on risk-based assessment of treatment effect heterogeneity in a randomized controlled trial. The aim of this study is to extend this approach to the observational setting using a standardized scalable framework. The proposed framework consists of five steps: 1) definition of the research aim, i.e., the population, the treatment, the comparator and the outcome(s) of interest; 2) identification of relevant databases; 3) development of a prediction model for the outcome(s) of interest; 4) estimation of relative and absolute treatment effect within strata of predicted risk, after adjusting for observed confounding; 5) presentation of the results. We demonstrate our framework by evaluating heterogeneity of the effect of thiazide or thiazide-like diuretics versus angiotensin-converting enzyme inhibitors on three efficacy and nine safety outcomes across three observational databases. We provide a publicly available R software package for applying this framework to any database mapped to the Observational Medical Outcomes Partnership Common Data Model. In our demonstration, patients at low risk of acute myocardial infarction receive negligible absolute benefits for all three efficacy outcomes, though they are more pronounced in the highest risk group, especially for acute myocardial infarction. Our framework allows for the evaluation of differential treatment effects across risk strata, which offers the opportunity to consider the benefit-harm trade-off between alternative treatments.

Introduction

Treatment effects often vary substantially across individual patients, causing overall effect estimates to be inaccurate for a significant proportion of the patients at hand[5,7]. Understanding this heterogeneity of treatment effects has been crucial for both personalized (or precision) medicine and comparative effectiveness research, giving rise to a wide range of approaches for its discovery, evaluation and application in clinical practice. A common approach to evaluating heterogeneity of treatment effect in clinical trials is through subgroup analyses. However, as these analyses are rarely adequately powered, they can lead to false conclusions of absence of heterogeneiy of treatment effect or exaggerate its presence[14,45]. In addition, patients differ in multiple characteristics simultaneously, resulting in much richer heterogeneit of treatment effect compared to the heterogeneity explored with regular one-variable-at-a-time subgroup analyses.

Baseline risk is a summary score inherently related to treatment effect that can be used to represent the variability in patient characteristics[13,14,38,97,98]. For example, an invasive coronary procedure—compared to medical treatment—improves survival in patients with myocardial infarction at high (predicted) baseline risk but not in those at low baseline risk[99]. It has also been shown that high-risk patients with pre-diabetes benefit substantially more from a lifestyle modification program than low-risk patients[49].

The recently proposed Predictive Approaches to Treatment effect Heterogeneity (PATH) statement provides systematic guidance on the application of risk-based methods for the assessment of heterogeneity of treatment effect in randomized controlled trial (RCT) data[16,17]. After risk-stratifying patients using an existing or an internally derived prediction model, risk stratum-specific estimates of relative and absolute treatment effect are evaluated. Several methods for predictive analysis heterogeneity of treatment effect have been adapted for use in observational data, but risk-based methods are still not readily available and have been highlighted as an important future research need[17].

The Observational Health Data Science and Informatics (OHDSI) collaborative has established a global network of data partners and researchers that aim to bring out the value of health data through large-scale analytics by mapping local databases to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)[31,100]. A standardized framework applying current best practices for comparative effectiveness studies within the OHDSI setting has been proposed[28]. This framework was successfully implemented in the Large-scale Evidence Generation and Evaluation across a Network of Databases for Hypertension (LEGEND-HTN) study. In this study, average

effects of all first-line hypertension treatment classes were estimated for a total of 55 outcomes across a global network of nine observational databases[33].

LEGEND-HTN found benefit for patients treated with thiazide or thiazide-like diuretics compared to angiotensin converting enzyme (ACE) inhibitors in terms of three main outcomes of interest, i.e., acute myocardial infarction (MI), hospitalization with heart failure, and stroke. Thiazide or thiazide-like diuretics also had a better safety profile compared to ACE inhibitors which, according to that study, makes them an attractive option for first-line treatment of hypertension. However, as already pointed out, overall (average) effect estimates may not be applicable to large portions of the target population due to strong variability of important patient characteristics. A risk-based analysis of treatment effect heterogeneity can add further insights to the results of LEGEND-HTN, both in understanding how treatment effects evolve with increasing baseline outcome risk and in identifying patient subgroups which could be targeted with a certain treatment.

Hereto, we focus on the three main outcomes of LEGEND-HTN (acute MI, hospitalization with heart failure, and stroke) and nine safety outcomes (hyponatremia, hypotension, acute renal failure, angioedema, kidney disease, cough, hyperkalemia, hypokalemia, and gastrointestinal bleeding). For our analyses, we develop a systematic framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases, extending the existing methodology from the RCT setting. The suggested framework is also implemented in an open-source, publicly available R-package. It is highly scalable and can be easily implemented across a network of observational databases mapped to OMOP-CDM.

Results

Overview

The proposed framework defines 5 distinct steps: 1) definition of the research aim; 2) identification of the databases within which the analyses will be performed; 3) prediction of outcomes of interest; 4) estimation of absolute and relative treatment effects within risk strata; 5) presentation of the results. We developed an open-source R-package for the implementation of the proposed framework and made it publicly available (https://github.com/OHDSI/RiskStratifiedEstimation). An overview of the entire framework can be found in ***Figure 1***.

As a demonstration, we evaluated treatment effect heterogeneity of thiazide or thiazide-like diuretics compared to ACE inhibitors using acute MI risk quarter-specific effect estimates, both on the relative and on the absolute scale. We

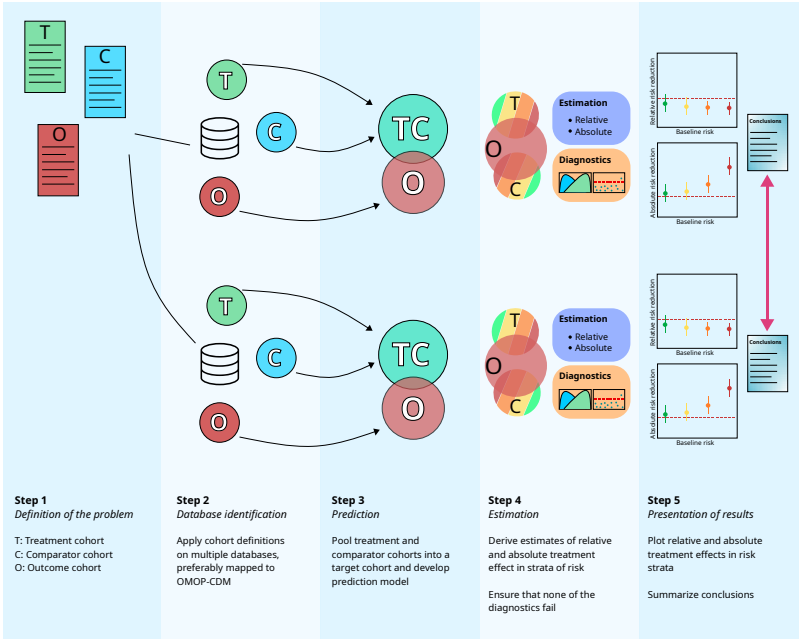

Figure 2.1: Illustration of the framework's application on two observational databases, preferably mapped to OMOP-CDM.

focused on three efficacy outcomes (acute MI, hospitalization with heart failure, and ischemic or hemorrhagic stroke) and nine safety outcomes (acute renal failure, kidney disease, cough, hyperkalemia, hypokalemia, gastrointestinal bleeding, hyponatremia, hypotension, and angioedema). We used data from three US-based claims databases.

Step 1: General definition of the research aim

We considered the following research aim: "compare the effect of thiazide or thiazide-like diuretics (T) to the effect of ACE inhibitors (C) in patients with established hypertension with respect to 12 outcomes $(O_1, \ldots, O_{12})$". The required cohorts are:

- Treatment cohort: Patients receiving any drug within the class of thiazide or thiazide-like diuretics with at least one year of follow-up before treatment initiation and a recorded hypertension diagnosis within that year.
- Comparator cohort: Patients receiving any drug within the ACE inhibitor class with at least one year of follow-up before treatment initiation and a recorded hypertension diagnosis within that year.

- Outcome cohorts: We considered three efficacy and nine safety outcome cohorts. These were patients in the database with a diagnosis of: acute MI; hospitalization with heart failure; ischemic or hemorrhagic stroke (efficacy outcomes); acute renal failure; kidney disease; cough; hyperkalemia; hypokalemia; gastrointestinal bleeding; hyponatremia; hypotension; angioedema (safety outcomes).

All cohort definitions were identical to the ones used in the multinational LEGEND-HTN study 16. More information can be found in the Supplementary Results (***Sections A and B***) and ***Supplementary Tables 1-19***.

Step 2: Identification of the databases

For our demonstration we used data from three US claims databases, namely IBM® MarketScan® Commercial Claims and Encounters (CCAE), IBM® MarketScan® Multi-State Medicaid (MDCD), and IBM® MarketScan® Medicare Supplemental Beneficiaries (MDCR). More information on the included databases can be found in Supplementary Results Section D. Our analyses included a total of 355,826 (CCAE), 54,835 (MDCD), and 37,882 (MDCR) patients initiating treatment with thiazide or thiazide-like diuretics and 930,629 (CCAE), 106,492 (MDCD), and 105,852 (MDCR) patients initiating treatment with ACE inhibitors (***Table 1***). Patient characteristics are available in ***Supplementary Tables 20-22***. Adequate numbers of patients were included in all strata of predicted acute MI risk (***Supplementary Table 23***).

Step 3: Prediction

We internally developed separate prediction models for 2-year acute MI risk in each of the three databases. The prediction models were fitted on the propensity score matched (1:1) subset of the entire study population, using a caliper of 0.2 and after excluding patients having the outcome at any time prior to treatment initiation. We considered a large set of candidate predictors containing patients' demographic information (age, sex), disease and medication history, and the Charlson comorbidity index (Romano adaptation) measured in the year prior to treatment initiation. As all three databases are mapped to OMOP-CDM, coding of all predictors was uniform across databases. This enables the development of the prediction models for acute MI risk in a uniform fashion across databases. However, due to the differences in data capture among databases, we cannot expect that all covariates will be present in all databases. We developed the prediction models using LASSO logistic regression with 3-fold cross validation for hyper-parameter selection. In ***Supplementary Table 24*** we show the available sample sizes on which the prediction models were developed, while in ***Supplementary Tables 25-27*** we

show the 20 selected covariates with the largest coefficients in each database. The models had moderate discriminative ability (internally validated) in CCAE and MDCD and lower discriminative ability in MDCR (***Table 2***).

Step 4: Estimation

In each database, we used patient-level predictions of the internally derived acute MI risk prediction model to stratify the patients into three acute MI risk groups RG-1, RG-2, and RG-3 (patients below 1% risk, patients between 1% and 1.5% risk, and patients above 1.5% risk). Within risk groups, in order to account for observed confounding, we further stratified the patients into five propensity score strata. Propensity score models were developed within each risk group separately using the same approach as in step 3 (LASSO logistic regression with a large set of predefined covariates). Risk group-specific relative treatment effects were estimated by averaging over the hazard ratio estimates derived from Cox regression models fitted in each propensity score stratum. Similarly, risk group-specific absolute treatment effects were estimated by averaging over the differences in Kaplan-Meier estimates in each propensity score stratum at 2 years after treatment initiation.

In all databases we found adequate overlap of the propensity score distributions across the risk groups, except for high-risk patients in CCAE (acute MI risk above 1.5%). Hence, the propensity scores should be able to adjust for observed confounding, except for high-risk CCAE patients (***Figure 2***). The covariate balance plots comparing covariate standardized mean differences before and after adjustment with the propensity scores confirmed strong imbalances for CCAE patients with acute MI predicted risk above 1.5% (***Figure 3***). Due to very limited overlap of the preference score distributions (***Figure 2***) and persisting imbalances after stratification on the propensity scores (***Figure 3***), we do not present the results for patients at risk above 1.5% for acute MI in CCAE. Additionally, a small number of characteristics remained slightly imbalanced even after stratification on the propensity scores for the two lower acute MI risk groups of MDCD (***Figure 3***). Therefore, results from analyses in this database should be interpreted with caution.

Finally, the distribution of the estimated relative risks with regard to a total of 76 negative control outcomes (Supplementary Results Section C) showed no evidence of residual confounding, except for CCAE (***Figure 4***) 17–19. Hazard ratios for CCAE (***Figure 4, Panel a***) were often significantly larger than 1 (true effect size). This suggests significant negative effects of thiazide or thiazide-like diuretics compared to ACE inhibitors on causally unrelated outcomes, indicating unresolved differences between the two treatment arms. Therefore, results from CCAE should be interpreted with caution, as resid-

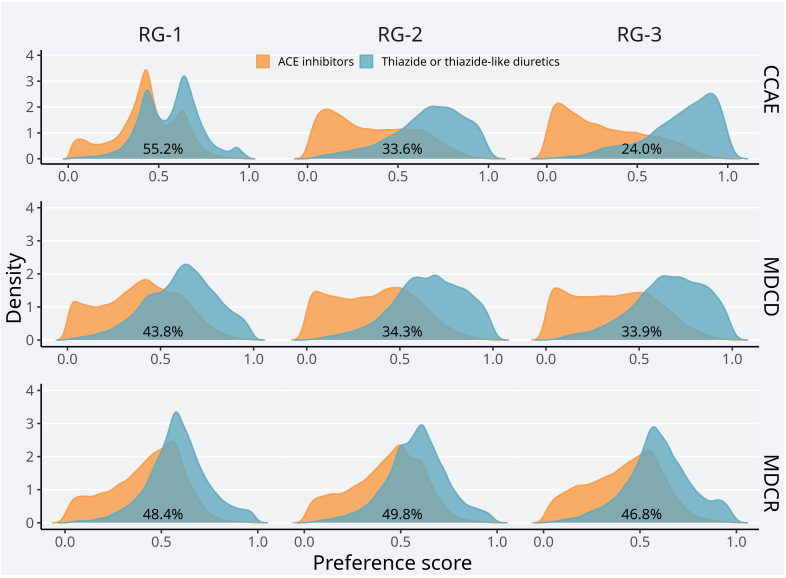

Figure 2.2: Preference score distributions within strata of predicted acute MI risk. RG-1 represents patients with acute MI risk lower than 1%; RG2 represents patients with acute MI risk between 1% and 1.5%; RG-3 represents patients with acute MI risk larger than 1.5%. The preference score is a transformation of the propensity score that adjusts for prevalence differences between populations. The percentages in each figure represent the amount of preference score overlap between treatment arms. Higher overlap of the preference score distributions indicates that patients in the target and the comparator cohorts are more similar in terms of the predicted probability of receiving treatment (thiazide or thiazide-like diuretics).

ual confounding may still be present, despite our propensity score adjustment. The results of the risk-stratified negative control analyses for each database can be found in Supplementary ***Figures 1-3***.

Step 5: Presentation of results

On average, thiazide or thiazide-like diuretics were beneficial compared to ACE inhibitors for all outcomes, except for hospitalization with heart failure in CCAE and stroke in MDCD (***Table 3***). The hazard ratios are in line with, but not equal to, those reported in the LEGEND-HTN study, mainly because of restricting time at risk to two years.

For the primary outcomes (acute MI, hospitalization with heart failure and stroke) relative treatment effect estimates of thiazide or thiazide-like diuretics versus ACE inhibitors varied substantially across risk groups, but no clear trends indicating an association between risk and relative treatment effect

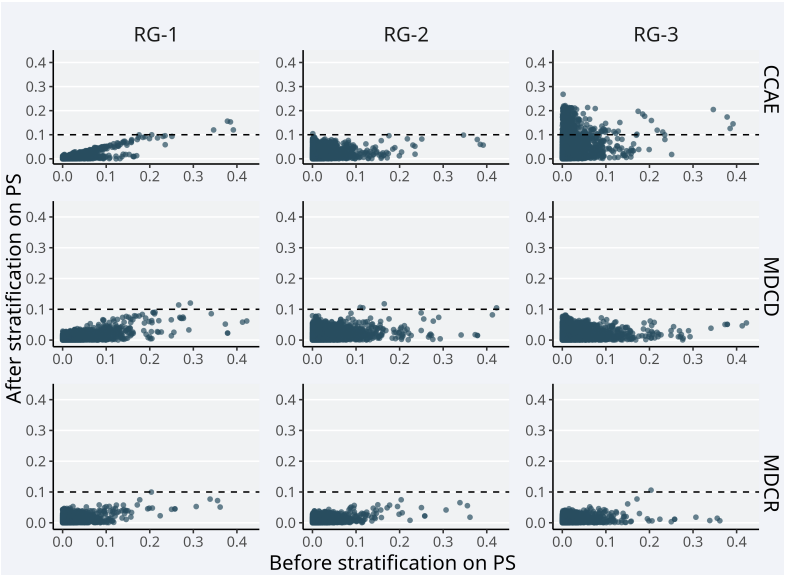

Figure 2.3: Patient characteristic balance for thiazide or thiazide-like diuretics and ACE inhibitors before and after stratification on the propensity scores. RG-1 represents patients with acute MI risk lower than 1%; RG-2 represents patients with acute MI risk between 1% and 1.5%; RG-3 represents patients with acute MI risk larger than 1.5%. Each point represents the standardized difference of means for a single covariate before (x-axis) and after (y-axis) stratification. A commonly used rule of thumb suggests that standardized mean differences above 0.1 after propensity score adjustment indicate insufficient covariate balance.

estimates were observed (***Figure 5***).

For acute MI, hazard ratios showed an increasing trend with increasing baseline acute MI risk in MDCD and CCAE, implying larger benefit on the relative scale for patients in the lower risk groups. This was less pronounced in MDCR (***Figure 5; Panel a***). For hospitalization with heart failure, hazard ratios were similar across all acute MI risk strata in MDCD, with a slightly decreasing trend favoring thiazide or thiazide-like diuretics (***Figure 5; Panel b***). In MDCR, these hazard ratios were very similar to MDCD for patients at acute MI risk higher than 1%. For patients below 1% acute MI risk, hazard ratios were close to 1 (negligible relative treatment effects) in all three databases. Finally, for stroke, the hazard ratios indicated a beneficial effect of thiazide or thiazide-like diuretics in all databases, but we found no clear trends in hazard ratios across acute MI risk groups (***Figure 5; Panel c***).

Absolute treatment effects (risk reduction) for acute MI and hospitalization

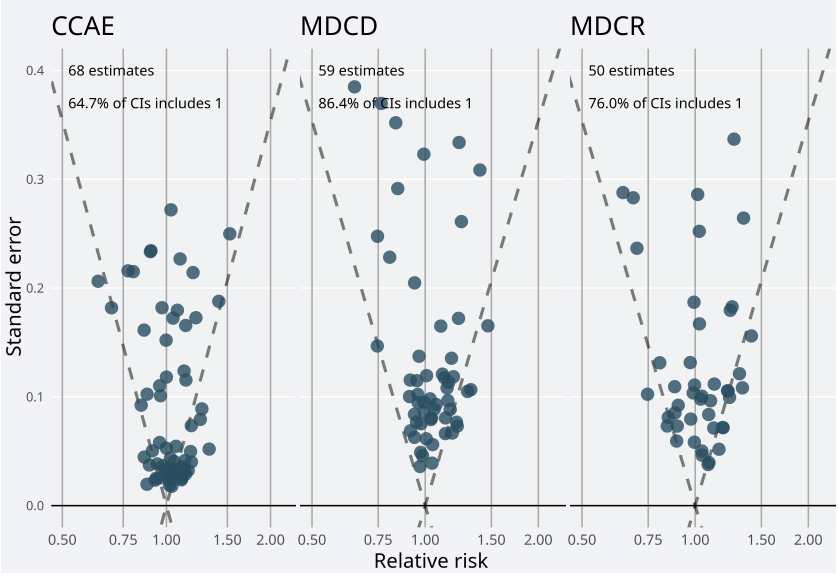

Figure 2.4: Effect size estimates for the negative controls (true hazard ratio = 1) in a CCAE, b MDCD, and c MDCR databases. Estimates below the diagonal dashed lines are statistically significant (different from the true effect size; alpha = 0.05). A well-calibrated estimator should include the true effect size within the 95% confidence interval, 95% of times

with heart failure tended to increase with increasing acute MI risk (***Figure 6; Panels a and b***). This was most evident in MDCD, where the absolute benefits for acute MI were 0.25% (0.03% to 0.48%; 95% CI) and 1.57% (0.49% to 2.65%; 95% CI) in the lowest and the highest acute MI risk group, respectively. Similarly, in MDCR these absolute benefits were -0.04% (-0.40% to 0.32%; 95% CI) and 0.70% (0.04% to 1.37%; 95% CI), respectively. For hospitalization with heart failure, these absolute benefits were -0.07% (-0.50% to 0.36%; 95% CI) and 2.31% (0.22% to 4.39%; 95% CI), respectively, in MDCD and -0.05% (-0.59% to 0.49%; 95% CI) and 0.97% (-0.16% to 2.09%; 95% CI), respectively, in MDCR. In CCAE, we found negligible treatment effects on the absolute scale for all three outcomes. Finally, for stroke, the differences on the absolute scale were small in all risk groups and databases (***Figure 6; Panel c***).

Across all databases and all risk groups (***Figure 7***), thiazide or thiazide-like diuretics reduced the risk for angioedema, cough, hyperkalemia, and hypotension, but were associated with increased risk of hypokalemia and hyponatremia. For cough and hypokalemia, the relative treatment effect tended to decrease

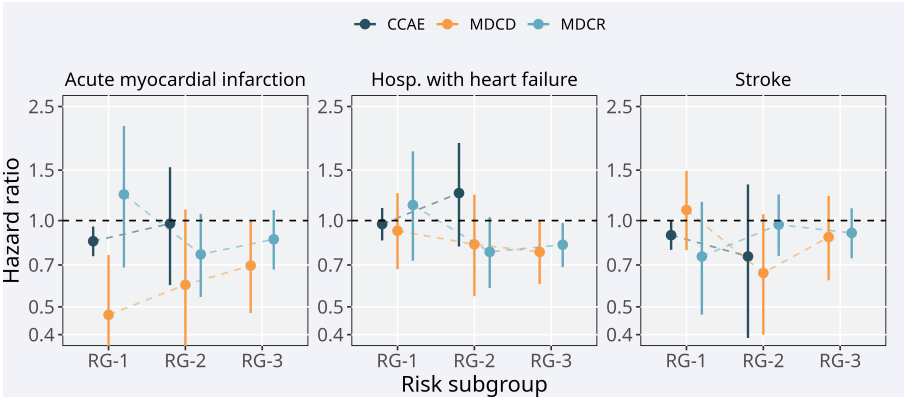

Figure 2.5: Relative treatment effects for main outcomes. Treatment effect heterogeneity for the main outcomes on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics compared to ACE inhibitors within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Hazard ratios estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors.

with increasing MI risk (hazard ratios moving closer to 1).

The absolute benefit for angioedema of thiazide or thiazide-like diuretics was negligible, despite the large treatment effect estimated on the relative scale (***Figure 8; Panel b***). The absolute risk increase of hypokalemia was large with thiazide or thiazide diuretics—as expected based on the effect estimates on the relative scale—across all risk strata (***Figure 8; Panel f***). This effect remained relatively constant across acute MI risk groups in MDCR, fluctuating between -4.13% and -3.25%. Similar effects on the absolute scale were observed in CCAE, where effect estimates were close to -5% for all patients below 1.5% risk of acute MI. A much larger hypokalemia risk increase with thiazide or thiazide-like diuretics was observed in MDCD, where the absolute effect estimates evolved from -9.89% (-11.23% to -8.54%; 95% CI) in patients below 1% acute MI risk to -15.58% (-23.78% to -7.38%; 95% CI) in patients above

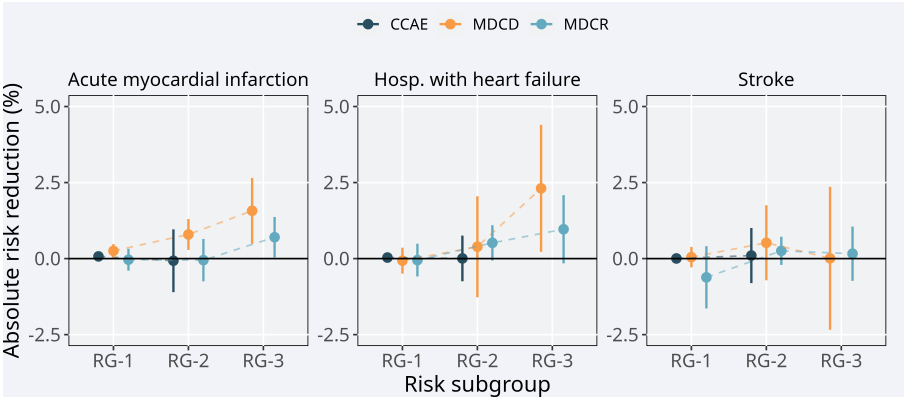

Figure 2.6: Absolute treatment effects for main outcomes. Treatment effect heterogeneity for the main outcomes on the absolute scale of thiazide or thiazide-like diuretics compared to ACE inhibitors within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5. Absolute treatment effects estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors.

1.5% acute MI risk. The absolute benefit estimates of thiazide or thiazide-like diuretics for cough ranged between 3.05% and 3.77% in CCAE, and between 2.32% and 3.73% in MDCR (***Figure 8; Panel c***). In MDCD, we observed a small risk increase of cough with thiazide or thiazide-like diuretics in patients at high acute MI baseline risk (-1.82% with a 95% CI from -7.82% to 4.17%). Finally, we observed a small risk increase of hyponatremia with thiazide or thiazide diuretics, which was more substantial in patients with high acute MI risk in MDCR (-1.91% with a 95% CI from -3.43% to -0.38%).

Interpretation

The overall benefits of thiazide or thiazide-like diuretics compared to ACE inhibitors that were observed in MDCR, in terms of acute MI and hospitalization with heart failure, were mainly driven by patients with predicted acute MI risk above 1.5%. Even in MDCD, where benefit on the absolute scale was

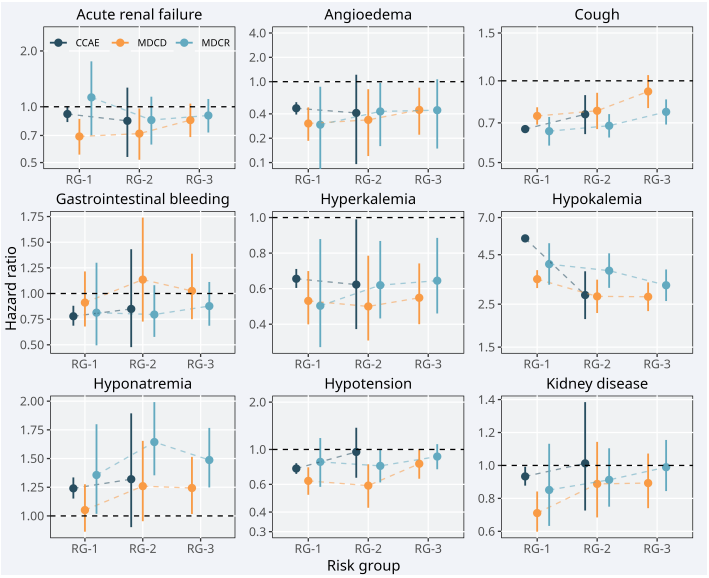

Figure 2.7: Relative treatment effects for safety outcomes. Treatment effect heterogeneity for the safety outcomes on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics compared to ACE inhibitors within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Hazard ratios estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. Bars represent 95% confidence intervals. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors.

observed across all acute MI risk strata, treatment effects were much larger in patients with predicted acute MI risk above 1.5%. In CCAE, where the majority of the patients had a predicted acute MI risk below 1%, we found negligible treatment effects. This provides further support for the similarity of the effect of thiazide or thiazide-like diuretics compared to ACE inhibitors in patients at low risk of acute MI.

Even though LEGEND-HTN found beneficial effects of thiazide or thiazide-like diuretics over ACE inhibitors in terms of several safety outcomes, there are still safety concerns when prescribing thiazide or thiazide-like diuretics. The hypokalemia and hyponatremia risk increase with thiazide or thiazide-like diuretics was not negligible in any of the acute MI risk strata. On the other hand, ACE inhibitor-related cough risk increase was also present in all databases and acute MI risk groups. Provided that absolute benefits of thiazide

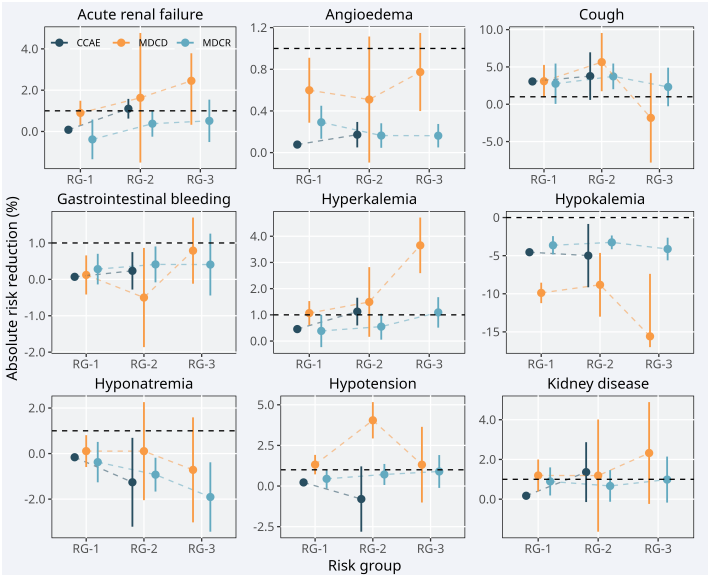

Figure 2.8: Absolute treatment effects for safety outcomes. Treatment effect heterogeneity for the safety outcomes on the absolute scale of thiazide or thiazide-like diuretics compared to ACE inhibitors within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Absolute treatment effects estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors.

or thiazide-like diuretics for the main outcomes (acute MI, hospitalization with heart failure, and stroke) were mainly observed in patients at high acute MI risk, the prescribing physician has to carefully weigh benefits and harms for individual patients.

Note that any conclusions drawn are for demonstration purposes only and should be interpreted under this very limited setting.

Sensitivity analyses

As a sensitivity analysis, we evaluated treatment effect heterogeneity of thiazide or thiazide-like diuretics compared to ACE inhibitors in patients with or without prior cardiovascular disease. We defined the set of patients without prior cardiovascular disease as the patients that had no occurrence in their medical history of any of the following conditions: heart valve disorder or transplanted heart valve, coronary artery disease, cardiac dysfunction, heart block,

unstable angina, atrial fibrillation, myocardial infarction, ventricular arrhythmia or cardiac arrest, ischemic heart disease, myocarditis or pericarditis, cardiomyopathy, cardiomegaly, heart failure, or stroke (ischemic or hemorrhagic). If patients had any of these conditions recorded in their medical history, they were assigned to the group with prior cardiovascular disease. We repeated our analyses using the exact same settings for both groups of patients.

In patients without prior cardiovascular disease, the estimates of the relative effect of thiazide or thiazide-like diuretics compared to ACE inhibitors on acute MI were similar to the original analyses—hazard ratios 0.90 (0.79 to 1.02; 95% CI), 0.52 (0.36 to 0.74; 95% CI), and 0.83 (0.65 to 1.05; 95% CI) in CCAE, MDCD, and MDCR respectively. In patients with prior cardiovascular disease the effect of thiazide or thiazide-like diuretics was stronger in CCAE—hazard ratio 0.73 (0.55 to 0.95; 95% CI)—but weaker in MDCD and MDCR—hazard ratios 0.78 (0.51 to 1.16; 95% CI) and 0.88 (0.66 to 1.15; 95% CI), respectively. In both sets of sensitivity analyses, risk stratified results showed trends comparable to the original analysis (***Supplementary Figures 4-11***).

Discussion

In this study we develop a risk-based framework for the assessment of treatment effect heterogeneity in large observational databases. Our framework fills a gap identified in the literature after the development of guidelines for performing such analyses in the RCT setting[16,17]. As an additional contribution we provide the software for implementing this framework in practice and make it publicly available. We made our software compatible to databases mapped to OMOP-CDM which allows researchers to easily implement our framework in a global network of healthcare databases. In our case study we demonstrate the use of our framework for the evaluation of treatment effect heterogeneity of thiazide or thiazide-like diuretics compared to ACE inhibitors on three efficacy and nine safety outcomes. We propose that this framework is implemented any time treatment effect estimation in high-dimensional observational data is undertaken.

In recent years, several methods for the analysis of treatment effect heterogeneity have been developed in the RCT setting[101]. However, low power and restricted prior knowledge on the mechanisms of variation in treatment effect are often inherent in RCTs, which are usually adequately powered only for the analysis of the primary outcome. Observational databases contain a large amount of information on treatment assignment and outcomes of interest, while also capturing key patient characteristics. They contain readily available data on patient sub-populations of interest on which no RCT has focused before either due to logistical or ethical reasons. However, observational

databases can be susceptible to biases, poorly measured outcomes and missingness, which may obscure true heterogeneity of treatment effect or falsely indicate it when there is none[39]. Therefore, inferences on both overall treatment effect estimates and heterogeneity of treatment effect need to rely on strong—often unverifiable—assumptions, despite the advancements and guidance on best practices. When evaluating treatment effect heterogeneity using a risk-based approach these issues may be compounded, mainly because of the risk of conflating confounding and effect modification. Well-designed observational studies on average replicate RCT results, even though often differences in magnitude may occur[102]. Our framework is in line with the recently suggested paradigm of high-throughput observational studies using consistent and standardized methods for improving reproducibility in observational research[103]. However, more empirical research comparing analyses of observational data and RCTs is required to assess the conditions under which different approaches for evaluating treatment effect heterogeneity provide credible results. Our software package can help support this research.

Our framework highlights the scale dependency of heterogeneity of treatment effect and how it relates to baseline risk. Treatment effect is mathematically determined by baseline risk, if we assume a constant non-zero effect size[20]. Patients with low baseline risk can only experience minimal benefits, before their risk is reduced to zero. In contrast, high risk patients can potentially have much larger absolute benefits. This becomes evident when evaluating the safety of thiazide or thiazide-like diuretics on angioedema and cough, both adverse events linked to treatment with ACE inhibitors. For angioedema, the substantial relative risk increase with ACE inhibitors only translated in a small risk increase on the absolute scale due to the limited baseline angioedema risk. Conversely, despite the small relative cough risk increase of ACE inhibitors, the large baseline cough risk resulted in larger absolute risk differences, compared to the other considered outcomes.

For patients with comorbidities the Guidelines of the American College of Cardiology often recommend initiation of treatment with ACE inhibitors, e.g. for patients with stable ischemic heart disease or patients with preserved ejection fraction[104]. Since these are patients with more severe medical conditions there may be a potential interaction of baseline acute MI risk with the propensity of receiving a thiazide or a thiazide-like diuretic. We do not formally test for that interaction, however, we observed that with increasing acute MI baseline risk, the overlap of the propensity score distributions decreases and the propensity score distributions for each treatment arm become more skewed, especially in CCAE and MDCD (***Figure 2***). This could potentially result in unobserved confounding being present even after propensity score adjustment. Indeed, in

CCAE, negative control analyses showed evidence of residual confounding and therefore results should be interpreted with caution. In risk-stratified negative control analyses we observed more evidence of residual confounding in patients with higher acute MI risk, which was, however, not identified in the other two databases.

The application of our framework in the case study is for demonstration purposes and there are several limitations to its conclusions. First, risk groups defined in each database were not defined using a universal prediction model, but using internally developed prediction models in each database. Future research could explore model combination or transfer learning methods for the development of universal risk prediction models. Second, death could be a competing risk. We could expand our framework in the future to potentially support sub-distribution hazard ratios and cumulative incidence reductions. Third, we only used the databases readily available to us and not all the available databases mapped to OMOP-CDM. Therefore, the generalizability of our results still needs to be explored in future studies. These studies should also address the particular aspects of the databases at hand, such as their sampling frame, the completeness of the data they capture and many other aspects that were not assessed in our demonstration. Fourth, we did not correct for multiplicity when presenting the results. We are interested in presenting trends in the data rather than detecting specific subgroups with significant treatment effects. The implementation of our framework, however, generates all the relevant output required for a researcher to correct for multiple testing, if that is required.

In conclusion, the case study demonstrates the feasibility of our framework for risk-based assessment of treatment effect heterogeneity in large observational data. It is easily applicable and highly informative whenever treatment effect estimation in high-dimensional observational data is of interest.

Methods

Step 1: General definition of the research aim

The typical research aim is: "to compare the effect of treatment to a comparator treatment in patients with a disease with respect to outcomes $O_1, \ldots, O_n$".

We use a comparative cohort design. This means that at least three cohorts of patients need to be defined at this stage of the framework:

- A single treatment cohort (T) which includes patients with disease receiving the target treatment of interest.
- A single comparator cohort (C) which includes patients with disease receiving the comparator treatment.

- One or more outcome cohorts $(O_1, \ldots, O_n)$ that contain patients developing the outcomes of interest.

Step 2: Identification of the databases

Including in our analyses multiple databases representing the population of interest potentially increases the generalizability of results. Furthermore, the cohorts should preferably have adequate sample size with adequate follow-up time to ensure precise effect estimation, even within smaller risk strata. Other relevant issues such as the depth of data capture (the precision at which measurements, lab tests, conditions are recorded) and the reliability of data entry should also be considered.

In our analyses, we used data from IBM® MarketScan® Commercial Claims and Encounters (CCAE), IBM® MarketScan® Medicaid (MDCD), and IBM® MarketScan® Medicare Supplemental Beneficiaries (MDCR). The New England Institutional Review Board (IRB) has determined that studies conducted in these databases are exempt from study-specific IRB review, as these studies do not qualify as human subjects research.

Step 3: Prediction

For our risk-based approach to adequately evaluate treatment effect heterogeneity, a well performing prediction model assigning patient-level risk for the outcome of interest needs to be available, either from literature or internally developed from the data at hand. For internally developing a risk prediction model we adopt the standardized framework focused on observational data that ensures adherence to existing guidelines[105–107]. We use the derived prediction model to separate the patient population into risk strata, within which treatment effects on both the relative and the absolute scale will be assessed.

For the development of the risk prediction model, we first need to define a target cohort of patients, i.e., the set of patients on whom the prediction model will be developed. In our case, the target cohort is generated by pooling the already defined treatment and comparator cohorts. We develop the prediction model on the propensity score-matched (1:1) subset of the pooled sample to avoid differentially fitting between treatment arms, thus introducing spurious interactions with treatment[15,90]. We also need to define a set of patients that experience the outcome of interest, i.e., the outcome cohort. Finally, we need to decide the time frame within which the predictions will be carried out, i.e., the patients' time at risk. Subsequently, we can develop the prediction model.

It is important that the prediction models display good discriminative ability to ensure that risk-based subgroups are accurately defined. A performance overview of the derived prediction models including discrimination and cali-

bration both in the propensity score matched subset, the entire sample and separately for treated and comparator patients should also be reported.

Step 4: Estimation

We estimate treatment effects (both on the relative and the absolute scale) within risk strata defined using the prediction model of step 3. We often consider four risk strata, but fewer or more strata can be considered depending on the available power for accurately estimating stratum-specific treatment effects. Effect estimation may be focused on the difference in outcomes for a randomly selected person from the risk stratum (average treatment effect) or for a randomly selected person from the treatment cohort within the risk stratum receiving the treatment under study (average treatment effect on the treated).

Any appropriate method for the analysis of relative and absolute treatment effects can be considered, as long as the this is done consistently in all risk strata. Common statistical metrics are odds ratios or hazard ratios for relative scale estimates and differences in observed proportions or differences in Kaplan-Meier estimates for absolute scale estimates, depending on the problem at hand. We estimate propensity scores within risk strata which we then use to match patients from different treatment cohorts or to stratify them into groups with similar propensity scores or to weigh each patient's contribution to the estimation process[25].

Prior to analyzing results, it is crucial to ensure that all diagnostics are passed in all risk strata. The standard diagnostics we carry out include analysis of the overlap of propensity score distributions and calculation of standardized mean differences of the covariates before and after propensity score adjustment. Finally, we use effect estimates for a large set of negative control outcomes—i.e., outcomes known to not be related with any of the exposures under study—to evaluate the presence of residual confounding not accounted for by propensity score adjustment[27,30,103].

Step 5: Presentation of results

In the presence of a positive treatment effect and a well-discriminating prediction model we expect an increasing pattern of the differences in the absolute scale, even if treatment effects remain constant on the relative scale across risk strata. Due to this scale-dependence of treatment effect heterogeneity, results should be assessed both on the relative and the absolute scale.

	Thiazides or thiazide-like diuretics			Ace inhibitors		
Outcome	Patients	Person years	Outcomes	Patients	Person years	Outcomes
CCAE						
Acute MI	355,826	204,593	405	930,369	584,167	1,813
Hosp. with HF	355,528	204,451	389	930,629	584,541	1,492
Stroke	354,446	203,792	425	923,604	579,736	1,636
MDCD						
Acute MI	54,835	21,440	76	106,492	51,481	440
Hosp. with HF	54,354	21,290	212	105,005	50,878	835
Stroke	54,259	21,179	149	104,410	50,334	562
MDCR						
Acute MI	37,882	24,642	161	105,852	74,990	732
Hosp. with HF	37,617	24,509	277	105,134	74,654	1,196
Stroke	37,248	24,267	261	102,502	72,705	977

Table 2.1: Number of patients, person years and events for the three efficacy outcomes of the study across the three databases after excluding patients with prior outcomes.

Population	CCAE	MDCD	MDCR
Matched	0.73 (0.71, 0.74)	0.76 (0.73, 0.79)	0.65 (0.62, 0.68)
Treatment	0.73 (0.71, 0.75)	0.82 (0.77, 0.86)	0.66 (0.62, 0.70)
Comparator	0.70 (0.67, 0.71)	0.74 (0.71, 0.76)	0.66 (0.64, 0.68)
Entire population	0.71 (0.70, 0.72)	0.76 (0.74, 0.78)	0.66 (0.64, 0.68)

Table 2.2: Discriminative ability (c-statistic) of the derived prediction models for acute MI in the matched set (development set), the treatment cohort, the comparator cohort, and the entire population in CCAE, MDCD, and MDCR. Values in parentheses are cross-validated 95the propensity score matched subset in each database on which the prediction models were developed. Treatment population is the set of patients receiving thiazide or thiazide-like diuretics in each database, while comparator population is the set of patients receiving ACE inhibitors. Finally, entire population refers to the combined set of treatment and comparator patients.

Population	CCAE	MDCD	MDCR
Matched	0.73 (0.71, 0.74)	0.76 (0.73, 0.79)	0.65 (0.62, 0.68)
Treatment	0.73 (0.71, 0.75)	0.82 (0.77, 0.86)	0.66 (0.62, 0.70)
Comparator	0.70 (0.67, 0.71)	0.74 (0.71, 0.76)	0.66 (0.64, 0.68)
Entire population	0.71 (0.70, 0.72)	0.76 (0.74, 0.78)	0.66 (0.64, 0.68)

Table 2.3: Hazard ratio estimates for the overall treatment effect of thiazide or thiazide-like diuretics compared to ACE inhibitors. Values in brackets are 95% confindence intervals.

CHAPTER 3

Estimating individualized treatment effects from randomized controlled trials

Chapter based on *Rekkas, A., Rijnbeek, P.R., Kent, D.M. et al. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk-based approaches. BMC Med Res Methodol 23, 74 (2023). https://doi.org/10.1186/s12874-023-01889-6*

Abstract

Background: Baseline outcome risk can be an important determinant of absolute treatment benefit and has been used in guidelines for "personalizing" medical decisions. We compared easily applicable risk-based methods for optimal prediction of individualized treatment effects. Methods: We simulated RCT data using diverse assumptions for the average treatment effect, a baseline prognostic index of risk, the shape of its interaction with treatment (none, linear, quadratic or non-monotonic), and the magnitude of treatment-related harms (none or constant independent of the prognostic index). We predicted absolute benefit using: models with a constant relative treatment effect; stratification in quarters of the prognostic index; models including a linear interaction of treatment with the prognostic index; models including an interaction of treatment with a restricted cubic spline transformation of the prognostic index; an adaptive approach using Akaike's Information Criterion. We evaluated predictive performance using root mean squared error and measures of discrimination and calibration for benefit. Results: The linear-interaction model displayed optimal or close-to-optimal performance across many simulation scenarios with moderate sample size (N=4,250; ~785 events). The restricted cubic splines model was optimal for strong non-linear deviations from a constant treatment effect, particularly when sample size was larger (N=17,000). The adaptive approach also required larger sample sizes. These findings were illustrated in the GUSTO-I trial. Conclusions: An interaction between baseline risk and treatment assignment should be considered to improve treatment effect predictions.

Introduction

Predictive approaches to heterogeneity of treatment effects aim at the development of models predicting either individualized effects or which of two (or more) treatments is better for an individual with regard to a specific outcome of interest[39]. These predictive approaches include both regression and machine learning techniques and are the subject of active research[14,92,108,109]. In prior work, we divided regression-based methods for the evaluation of treatment effect heterogeneity in three broader categories: risk modeling, treatment effect modeling and optimal treatment regime methods[101]. Risk modeling methods use only prognostic factors to define patient subgroups, relying on the mathematical dependency between baseline risk and treatment effect[13,14]. Treatment effect modeling methods use both prognostic factors and treatment effect modifiers to explore characteristics that interact with the effects of therapy. They can be applied in one stage by directly modeling treatment-covariate interactions, in which case penalization of the interaction effects is needed to reduce the effects of overfitting[56], or in two stages that rely on updating working absolute benefit models[59,62]. Optimal treatment regime methods focus primarily on treatment effect modifiers in order to classify the trial population into those who benefit from treatment and those who do not[66,68–70].

In a previous simulation study, modeling treatment-covariate interactions often led to poorly calibrated predictions of benefit on the absolute scale (risk difference between treatment arms), compared to risk-modeling methods[90]. In the presence of true treatment-covariate interactions, however, effect modeling methods were better able to separate lower from higher benefit patients[90,110]. By assuming treatment effect is a function of baseline risk, risk modeling methods impose a restriction on the shape of treatment effect heterogeneity. With smaller sample sizes or limited information on effect modification, risk modeling methods, because of their reduced complexity, can provide a good option for evaluating treatment effect heterogeneity. Conversely, with larger sample sizes and/or a limited set of well-studied strong effect modifiers, treatment effect modeling methods can potentially result in a better bias-variance tradeoff. Therefore, the setting in which treatment effect heterogeneity is evaluated is crucial for the selection of the optimal approach.

Risk modeling methods predict similar treatment benefit for patients with similar baseline outcome risk, i.e. a similar probability of experiencing the outcome of interest in the absence of treatment. These methods are not new and are quite intuitive to practitioners[101]. Often medical guidelines rely on a risk stratified approach to target treatments to different patients. In addition, reanalyses of studies that only looked at overall results using risk stratification

often resulted to important insight on how treatment effects varied for different patients. For example, a risk stratified analysis of patients with acute myocardial infarction (MI) based on the Thrombolysis in Myocardial Infarction (TIMI) risk score found no benefit for patients who underwent primary angioplasty compared to fibrinolysis. However, there was a significant benefit for patients with a high TIMI score[99]. Infants at lower risk of bronchopulmonary dysplasia benefit relatively more from vitamin A therapy than infants at higher risk[111]. Finally, higher risk prediabetic patients benefit relatively more from metformin than lower risk patients[49].

Most often, risk-modeling approaches are carried out in two steps: first a risk prediction model is developed externally or internally on the entire RCT population, "blinded" to treatment; then the RCT population is stratified using this prediction model to evaluate risk-based treatment effect variation[13,16,17]. This approach identified substantial absolute treatment effect differences between low-risk and high-risk patients in a re-analysis of 32 large trials[47]. However, even though treatment effect estimates at the risk subgroup level may be accurate, these estimates may not apply to individual patients, as homogeneity of treatment effects is assumed within risk strata. With stronger overall treatment effect and larger variability in predicted risks, patients assigned to the same risk subgroup may still differ substantially with regard to their benefits from treatment.

In the current simulation study, we aim to summarize and compare different risk-based models for predicting treatment effects. We simulate different relations between baseline risk and treatment effects and also consider potential harms of treatment. We illustrate the different models by a case study of predicting individualized effects of treatment for acute myocardial infarction in a large RCT.

Methods

We observe RCT data (Z, X, Y), where for each patient $Z_i = 0, 1$ is the treatment status, $Y_i = 0, 1$ is the observed outcome and X_i is a set of measured covariates. Let $\{Y_i(z), z = 0, 1\}$ denote the unobservable potential outcomes. We observe $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. We are interested in predicting the conditional average treatment effect (CATE),

$$\tau(x) = E\{Y(0) - Y(1)|X = x\}$$

Assuming that $(Y(0), Y(1)) \perp\!\!\!\perp Z|X$, as we are in the RCT setting, we can predict CATE from

$$\begin{aligned}\tau(x) &= E\{Y(0) \mid X = x\} - E\{Y(1) \mid X = x\} \\ &= E\{Y \mid X = x, Z = 0\} - E\{Y \mid X = x, Z = 1\}\end{aligned}$$

Simulation scenarios

We simulated a typical RCT, comparing equally-sized treatment and control arms in terms of a binary outcome. For each patient we generated 8 baseline covariates $x_1, \ldots, x_4 \sim N(0, 1)$ and $x_5, \ldots, x_8 \sim B(1, 0.2)$. Outcomes in the control arm were generated from Bernoulli variables with true probabilities following a logistic regression model including all baseline covariates, i.e. $P(Y(0) = 1 \mid X = x) = \text{expit}(lp_0) = e^{lp_0}/(1 + e^{lp_0})$, with $lp_0 = lp_0(x) = x^t\beta$. In the base scenarios coefficient values β were such, that the control event rate was 20% and the discriminative ability of the true prediction model measured using Harrell's c-statistic was 0.75. The c-statistic represents the probability that for a randomly selected discordant pair from the sample (patients with different outcomes) the prediction model assigns larger risk to the patient with the worse outcome. For the simulations this was achieved by selecting β values such that the true prediction model would achieve a c-statistic of 0.75 in a simulated control arm with 1,000,000 patients. In the base case scenario we achieved that by setting $\beta = (-2.08, 0.49, \ldots, 0.49)^t$.

Outcomes in the treatment arm were first generated using 3 simple scenarios: absent (OR = 1), moderate (OR = 0.8) or strong (OR = 0.5) constant relative treatment effect. We then introduced linear, quadratic and non-monotonic deviations from constant treatment effects using:

$$lp_1 = \gamma_2(lp_0 - c)^2 + \gamma_1(lp_0 - c) + \gamma_0,$$

where lp_1 is the true linear predictor in the treatment arm, so that $P(Y(1) = 1 \mid X = x) = \text{expit}(lp_1)$, $\gamma = (\gamma_0, \gamma_1, \gamma_2)^t$ controls the base treatment effect and the shape of evolution of treatment effect as a function of baseline risk (type and strength of deviations from the constant treatment effect setting, while c allows us to shift the posited function to achieve the desired overall event rates. Finally, we incorporated constant absolute harms for all treated patients, such that $P(Y(1) = 1|X = x) = \text{expit}(lp_1) + \text{harm}$.

The sample size for the base scenarios was set to 4,250 (80% power for the detection of a marginal OR of 0.8 with the standard alpha of 5%). We evaluated the effect of smaller or larger sample sizes of 1,063 and 17,000, respectively. We also evaluated the effect of risk model discriminative ability, adjusting the baseline covariate coefficients, such that the AUC of the regression model in

the control arm was 0.65 and 0.85, respectively. These settings resulted in a simulation study of 648 scenarios covering the heterogeneity of treatment effects observed in 32 large trials as well as many other potential variations of risk-based treatment effect (Supplement, Sections 2 and 3)[47].

We analyzed the sensitivity of the results to correlation between baseline characteristics. We first sampled 8 continuous variables $W_1, \dots, W_8 \sim N(0, \Sigma)$. We then generated four continuous baseline covariates from $X_1 = W_1, \dots, X_4 = W_4$ and four binary covariates with 20% prevalence from $X_5 = I(W_5 > z_{0.8}), \dots, X_8 = I(W_8 > z_{0.8})$, where I is the indicator function and $P(U \leq 0.8) = z_{0.8}$ for random variable $U \sim N(0, 1)$. The covariance matrix Σ was such that $cor(X_i, X_j) = 0.5$ for any $i \neq j$. To ensure that the outcome rate in the untreated subset was 20% and that true prediction c-statistic remained equal to the nominal values of the main simulation analyses, we adjusted the coefficients of the true outcome model. More details on the sensitivity analyses can be found in the Supplement, section 9.

Individualized risk-based benefit predictions

In each simulation run, we internally developed a prediction model on the entire population, using a logistic regression model with main effects for all baseline covariates and treatment assignment. Individual risk predictions were derived by setting treatment assignment to 0. A more intuitive approach would be to derive the prediction model solely on the control patients. However, this has been shown to lead to biased benefit predictions, because with limited sample size the model will be overfitted to the control arm and induce spurious treatment interactions[15,18,90].

We compared different methods for predicting absolute treatment benefit, that is the risk difference between distinct treatment assignments. We use the term absolute treatment benefit to distinguish from relative treatment benefit that relies on the ratio of predicted risk under different treatment assignments.

A stratified heterogeneity of treatment effect method has been suggested as an alternative to traditional subgroup analyses[16,17]. Patients are stratified into equally-sized risk strata—in this case based on risk quartiles. Absolute treatment effects, within risk strata, expressed as absolute risk differences, are estimated by the difference in event rate between control and treatment arm patients. We considered this approach as a reference, expecting it to perform worse than the other candidates, as its objective is to provide an illustration of treatment effect heterogeneity rather than to optimize individualized benefit predictions.

Second, we fitted a logistic regression model which assumes constant relative treatment effect (constant odds ratio), that is, $P(Y = 1|X = x, Z = z; \widehat{\beta})$.

Hence, absolute benefit is predicted from $\tau(x;\hat{\beta}) = \text{expit}(\hat{lp}_0) - \text{expit}(\hat{lp}_0 + \delta_1)$, where δ_1 is the log of the assumed constant odds ratio and $\hat{lp}_0 = lp_0(x;\hat{\beta}) = x^t\hat{\beta}$ the linear predictor of the estimated baseline risk model.

Third, we fitted a logistic regression model including treatment, the risk linear predictor, and their linear interaction, that is, $P(Y = 1|X = x, Z = z;\hat{\beta}) = \text{expit}(\delta_0 + \delta_1 z + \delta_2\hat{lp}_0 + \delta_3 z\hat{lp}_0)$ Absolute benefit is then estimated from $\tau(x;\hat{\beta}) = \text{expit}(\delta_0 + \delta_2\hat{lp}_0) - \text{expit}\Big((\delta_0 + \delta_1) + (\delta_2 + \delta_3)\hat{lp}_0\Big)$. We will refer to this method as the linear interaction approach.

Fourth, we used restricted cubic splines (RCS) to relax the linearity assumption on the effect of the linear predictor[112]. We considered splines with 3 (RCS-3), 4 (RCS-4) and 5 (RCS-5) knots, together with their interaction with treatment, to compare models with different levels of flexibility (Supplement, section 4).

Finally, we considered an adaptive approach using Akaike's Information Criterion (AIC) for model selection. More specifically, we ranked the constant relative treatment effect model, the linear interaction model, and the RCS models with 3, 4, and 5 knots based on their AIC and selected the one with the lowest value. The extra degrees of freedom were 1 (linear interaction), 2, 3 and 4 (RCS models) for these increasingly complex interactions with the treatment effect.

Evaluation metrics

We evaluated the predictive accuracy of the considered methods by the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\tau(x_i) - \hat{\tau}(x_i)\right)^2}$$

We compared the discriminative ability of the methods under study using c-for-benefit and the integrated calibration index (ICI) for benefit (Supplement, section 6).

Since true patient-specific benefit is unobservable, we calculated observed benefit using the following approach: patients in each treatment arm are ranked based on their predicted benefit and then matched 1:1 on predicted benefit across treatment arms. Observed treatment benefit is defined as the difference of observed outcomes between the untreated and the treated patient of each matched patient pair. Since matching may not be perfect, that is, predicted benefits for the patients of the pair may not be equal, pair-specific predicted benefit is defined as the average of predicted benefit within each matched patient pair[22]. Then, the c-for-benefit represents the probability that from two randomly chosen predicted benefit-matched patient pairs with unequal

observed benefit, the pair with greater observed benefit also has a higher predicted benefit

We evaluated calibration in a similar manner, using the integrated calibration index (ICI) for benefit[113]. The observed benefits are regressed on the predicted benefits using a locally weighted scatterplot smoother (loess). The ICI-for-benefit is the average absolute difference between predicted and smooth observed benefit. Values closer to 0 represent better calibration.

For each scenario we performed 500 replications, within which all the considered models were fitted. We simulated a super-population of size 500,000 for each scenario within which we calculated RMSE and discrimination and calibration for benefit of all the models in each replication.

Empirical illustration

We demonstrated the different methods using 30,510 patients with acute myocardial infarction (MI) included in the GUSTO-I trial. 10,348 patients were randomized to tissue plasminogen activator (tPA) treatment and 20,162 were randomized to streptokinase. The outcome of interest was 30-day mortality (total of 2,128 events), recorded for all patients.

This dataset has been used extensively in prior studies[19,114]. Therefore, we used the same set of seven covariates that was previously used to fit a logistic regression model (age, Killip class, systolic blood pressure, heart rate, an indicator of previous MI, and the location of MI) along with a binary covariate for treatment indication, to predict 30-day mortality risk (Supplement, Section 10). Predicted baseline risk is derived by setting the treatment indicator to 0 for all patients.

Results

Simulation

The constant treatment effect approach outperformed other approaches in the base case scenario (N = 4,250; OR = 0.8; c-statistic= 0.75; no absolute treatment harm) with a true constant treatment effect (median RMSE: constant treatment effect 0.009; linear interaction 0.014; RCS-3 0.018). The linear interaction model was optimal under true linear deviations (median RMSE: constant treatment effect 0.027; linear interaction 0.015; RCS-3 0.018; Figure 1 panels A-C) and even in the presence of true quadratic deviations (median RMSE: constant treatment effect 0.057; linear interaction 0.020; RCS-3 0.021; Figure 1 panels A-C) from a constant relative treatment effect. With non-monotonic deviations, RCS-3 slightly outperformed the linear interaction model (median RMSE: linear interaction 0.019; RCS-3 0.018; Figure 1 panel

D). With strong treatment-related harms the results were very similar in most scenarios (Figure 1 panels A-C). Under non-monotonic deviations the optimal performance of RCS-3 was more pronounced (median RMSE: linear interaction 0.024; RCS-3 0.019; Figure1 panel D). A stronger average treatment effect (OR=0.5) resulted in higher variability of the true treatment effects on the absolute scale (difference in true outcome probabilities between treatment arms) and consequently to larger RMSE for all approaches. When we assumed a stronger relative treatment effect, the relative differences between approaches were similar to the base-case scenario (Supplement, Figure S10).

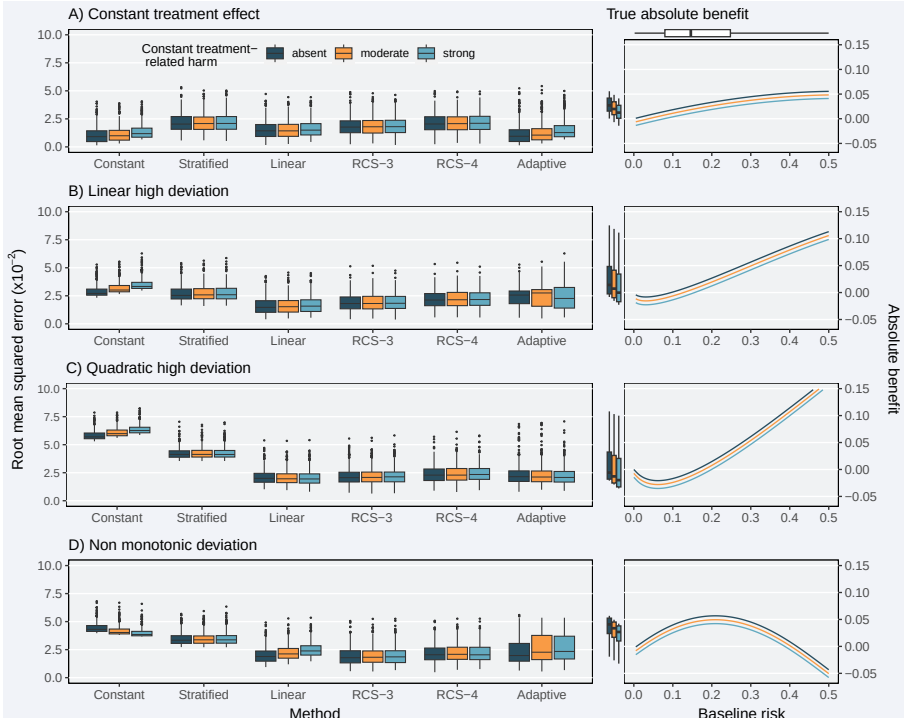

Figure 3.1: RMSE of the considered methods across 500 replications was calculated from a simulated super-population of size 500,000. The scenario with true constant relative treatment effect (panel A) had a true prediction c-statistic of 0.75 and sample size of 4250. The RMSE is also presented for strong linear (panel B), strong quadratic (panel C), and non-monotonic (panel D) deviations from constant relative treatment efects. Panels on the right side present the true relations between baseline risk (x-axis) and absolute treatment beneft (y-axis). The 2.5, 25, 50, 75, and 97.5 percentiles of the risk distribution are expressed by the boxplot on the top. The 2.5, 25, 50, 75, and 97.5 percentiles of the true beneft distributions are expressed by the boxplots on the side of the right-handside panel.

The adaptive approach had limited loss of performance in terms of the median RMSE to the best-performing method in each scenario. However, compared to the best-performing approach, its RMSE was more variable in scenarios with linear and non-monotonic deviations, especially when also including moderate or strong treatment-related harms. On closer inspection, we found that this behavior was caused by selecting the constant treatment effect model in a substantial proportion of the replications (Supplement, Figure S3).

Increasing the sample size to 17,000 favored RCS-3 the most (Figure 2). The difference in performance with the linear interaction approach was more limited in settings with a constant treatment effect (median RMSE: linear interaction 0.007; RCS-3 0.009) and with a true linear interaction (median RMSE: linear interaction 0.008; RCS-3 0.009) and more emphasized in settings with strong quadratic deviations (median RMSE: linear interaction 0.013; RCS-3 0.011) and non-monotonic deviations (median RMSE: linear interaction 0.014; RCS-3 0.010). Due to the large sample size, the RMSE of the adaptive approach was even more similar to the best-performing method, and the constant relative treatment effect model was less often wrongly selected (Supplement, Figure S4).

Similarly, when we increased the c-statistic of the true prediction model to 0.85 (OR = 0.8 and N = 4,250), RCS-3 had the lowest RMSE in the case of strong quadratic or non-monotonic deviations and very comparable performance to the – optimal – linear interaction model in the case of strong linear deviations (median RMSE of 0.016 for RCS-3 compared to 0.014 for the linear interaction model; Figure 3). Similar to the base case scenario the adaptive approach wrongly selected the constant treatment effect model (23% and 25% of the replications in the strong linear and non-monotonic deviation scenarios without treatment-related harms, respectively), leading to increased variability of the RMSE (Supplement, Figure S5).

With a true constant relative treatment effect, discrimination for benefit was only slightly lower for the linear interaction model, but substantially lower for the non-linear RCS approaches (Figure 4; panel A). With strong linear or quadratic deviations from a constant relative treatment effect, all methods discriminated quite similarly (Figure 4 panels B-C). With non-monotonic deviations, the constant effect model had much lower discriminative ability compared to all other methods (median c-for-benefit of 0.500 for the constant effects model, 0.528 for the linear interaction model and 0.530 Figure 4; panel D). The adaptive approach was unstable in terms of discrimination for benefit, especially with treatment-related harms. With increasing number of RCS knots, we observed decreasing median values and increasing variability of the c-for-benefit in all scenarios. When we increased the sample size to 17,000 we

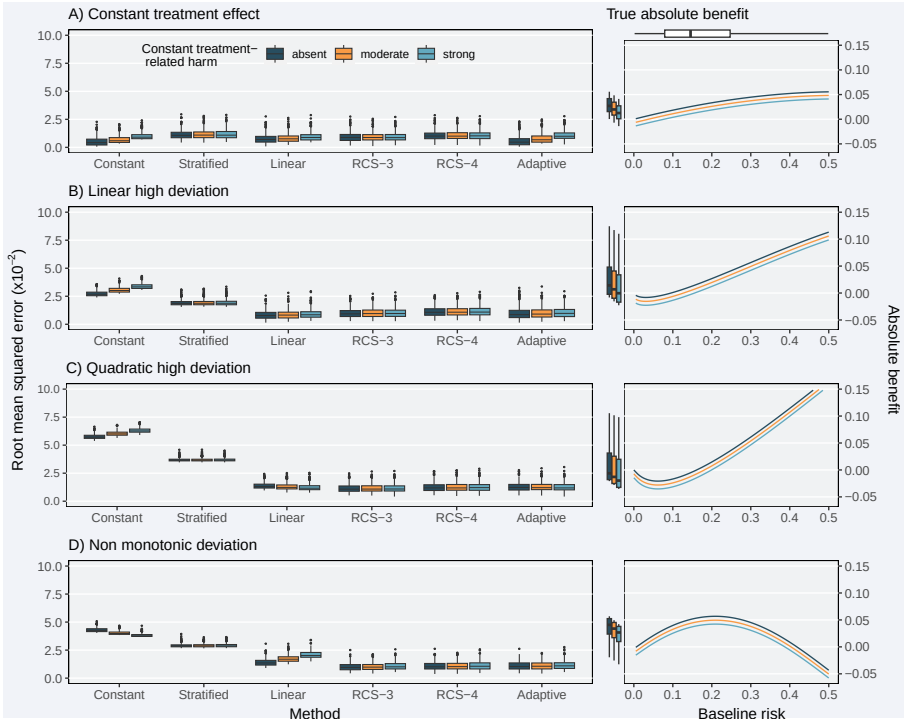

Figure 3.2: RMSE of the considered methods across 500 replications calculated in simulated samples of size 17,000 rather than 4,250 in Figure 3.1. RMSE was calculated on a super-population of size 500,000

observed similar trends, however the performance of all methods was more stable (Supplement, Figure S6). Finally, when we increased the true prediction c-statistic to 0.85 the adaptive approach was, again, more conservative, especially with non-monotonic deviations and null or moderate treatment-related harms (Supplement, Figure S7).

In terms of calibration for benefit, the constant effects model outperformed all other models in the scenario with true constant treatment effects, but was miscalibrated for all deviation scenarios (Figure 5). The linear interaction model showed best or close to best calibration across all scenarios and was only outperformed by RCS-3 in the case of non-monotonic deviations and treatment-related harms (Figure 5 panel D). The adaptive approach was worse calibrated under strong linear and non-monotonic deviations compared to the linear interaction model and RCS-3. When we increased the sample size to 17,000 (Supplement, Figure S8) or the true prediction c-statistic to 0.85 (Supplement, Figure S9), RCS-3 was somewhat better calibrated than the linear

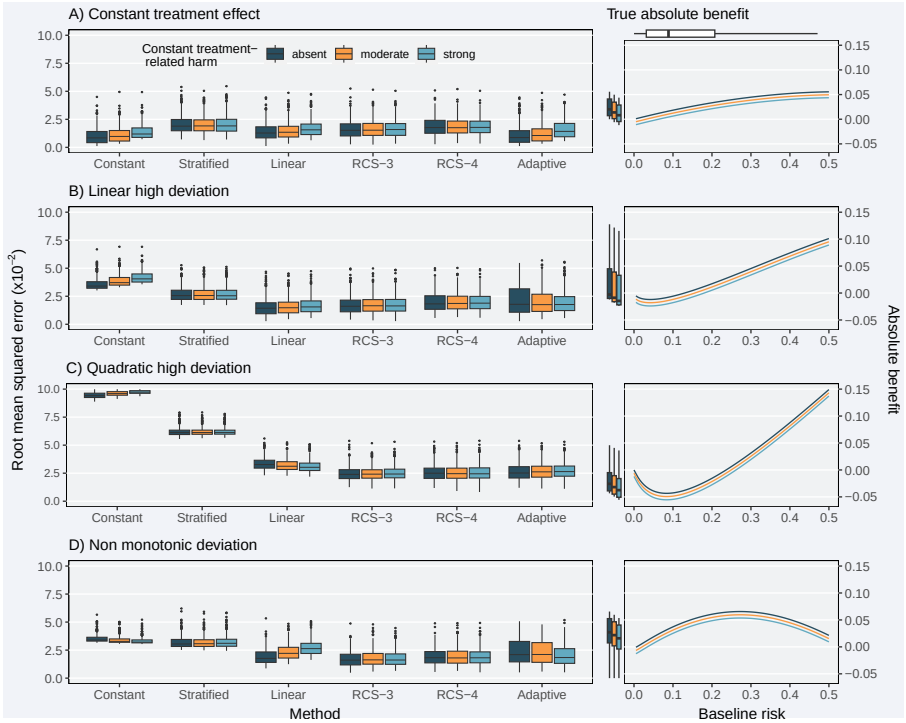

Figure 3.3: RMSE of the considered methods across 500 replications calculated in simulated samples 4,250. True prediction c-statistic of 0.85. RMSE was calculated on a super-population of size 500,000

interaction model with strong quadratic deviations.

Our main conclusions remained unchanged in the sensitivity analyses where correlations between baseline characteristics were introduced (Supplement, Figures S16, S17, and S18).

The results from all individual scenarios can be explored online at https://mi-erasmusmc.shinyapps.io/HteSimulationRCT/. Additionally, all the code for the simulations can be found at https://github.com/mi-erasmusmc/HteSimulationRCT.

Empirical illustration

We used the derived prognostic index to fit a constant treatment effect, a linear interaction and an RCS-3 model individualizing absolute benefit predictions. Following our simulation results, RCS-4 and RCS-5 models were excluded. Finally, an adaptive approach with the 3 candidate models was applied.

Predicted absolute benefit was derived as the difference of predicted acute MI

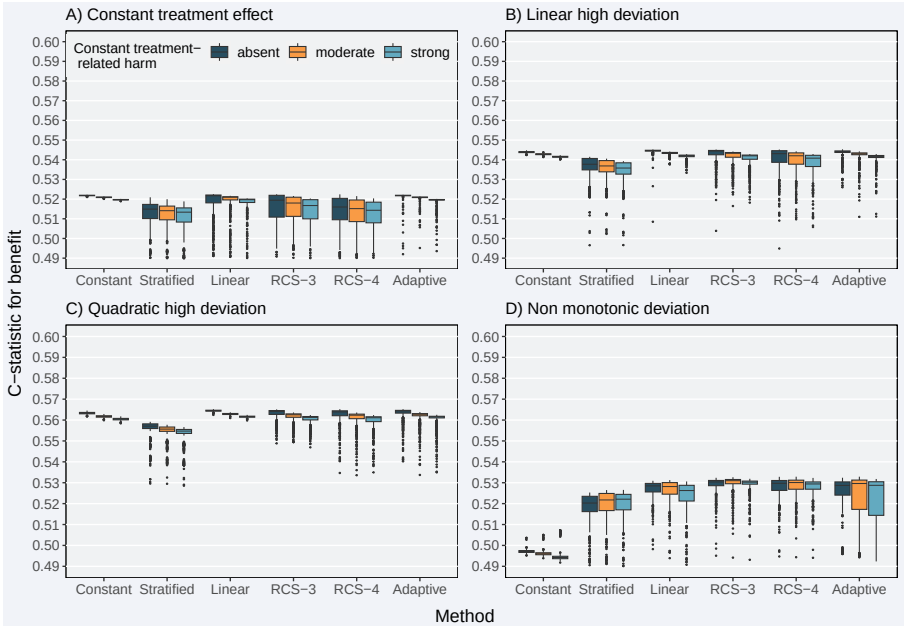

Figure 3.4: Discrimination for beneft of the considered methods across 500 replications calculated in simulated samples of size 4,250 using the c-statistic for beneft. The c-statistic for beneft represents the probability that from two randomly chosen matched patient pairs with unequal observed beneft, the pair with greater observed beneft also has a higher predicted beneft. True prediction c-statistic of 0.75.

risk between treatment arms, if all other predictors remained unchanged. All considered methods provided similar fits, predicting increasing absolute benefits for patients with higher baseline risk predictions, and followed the evolution of the stratified estimates closely (Figure 6). The constant treatment effect model had somewhat lower AIC compared to the linear interaction model (AIC: versus 9,342), equal cross-validated discrimination (c-for-benefit: 0.525), and slightly better cross-validated calibration (ICI-for benefit: 0.010 versus 0.012). In conclusion, although the sample size (30,510 patients; 2,128 events) allowed for flexible modeling approaches, a simpler constant treatment effect model is adequate for predicting absolute 30-day mortality benefits of treatment with tPA in patients with acute MI.

Discussion

The linear interaction and the RCS-3 models displayed very good performance under many of the considered simulation scenarios. The linear interaction model was optimal in cases with moderate sample sizes (4.250 patients; ~785

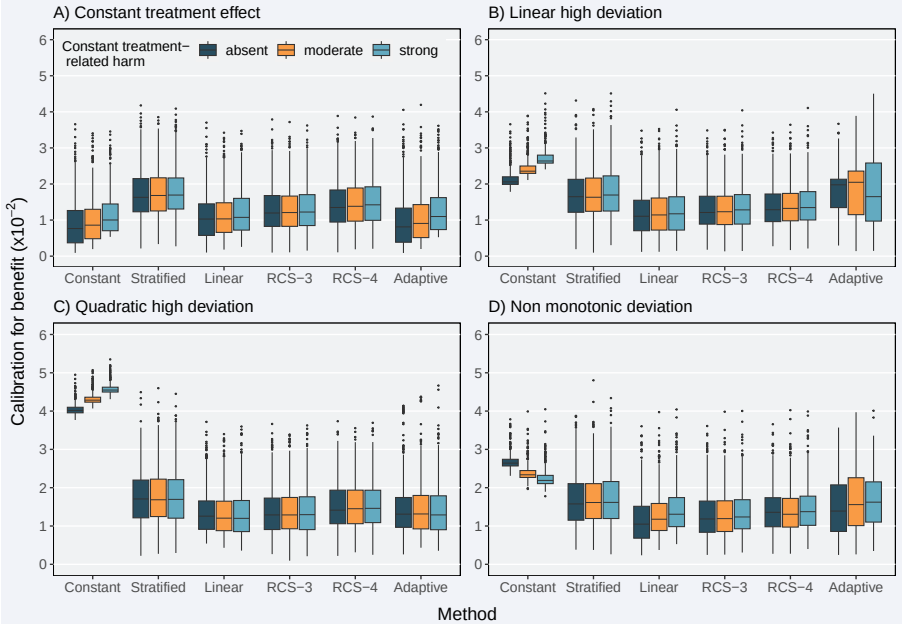

Figure 3.5: Calibration for beneft of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction c-statistic of 0.75 and sample size of 4,250.

events) and moderately performing baseline risk prediction models, that is, it had lower RMSE, was better calibrated for benefit and had better discrimination for benefit, even in scenarios with strong quadratic deviations. In scenarios with true non-monotonic deviations, the linear interaction model was outperformed by RCS-3, especially in the presence of treatment-related harms. Increasing the sample size or the prediction model's discriminative ability favored RCS-3, especially in scenarios with strong non-linear deviations from a constant treatment effect.

Our simulation results clearly express the trade-off between the advantages of flexibly modeling the relationship between baseline risk and treatment effect and the disadvantages of overfitting this relationship to the sample at hand. With infinite sample size, the more flexible approach (here RCS) will be optimal, but in practice, with limited sample size, parsimonious models may be preferable. Even with the substantial sample size of our base case scenario, the (less flexible) linear interaction model performed better than the (more flexible) RCS approach for most simulation settings. The even less flexible constant treatment effect model, however, was only optimal when the treatment effect

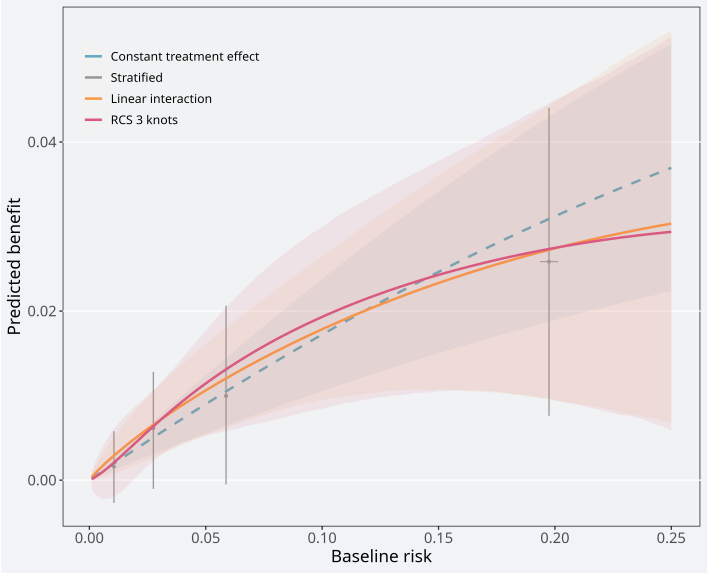

Figure 3.6: 6 Individualized absolute beneft predictions based on baseline risk when using a constant treatment efect approach, a linear interaction approach and RCS smoothing using 3 knots. Risk stratifed estimates of absolute beneft are presented within quartiles of baseline risk as reference. 95% confdence bands were generated using 10,000 bootstrap resamples, where the prediction model was reftted in each run to capture the uncertainty in baseline risk predictions. For the risk stratifcation approach, we also provide 95% confdence intervals for the baseline risk quarter-specifc average predicted risk over the 10,000 bootstrap samples.

was truly constant. Moreover, the assumption of a constant treatment effect may often be too strong[47,60].

RCS-4 and RCS-5 were too flexible in all considered scenarios, as indicated by higher RMSE, increased variability of discrimination for benefit and worse calibration of benefit predictions. Even with larger sample sizes and strong quadratic or non-monotonic deviations, these more flexible methods did not outperform the simpler RCS-3 approach. Higher flexibility may only be helpful under more extreme patterns of treatment effect heterogeneity compared to the quadratic deviations considered here. Considering interactions in RCS-3 models as the most complex approach often may be reasonable.

Our results can also be interpreted in terms of bias-variance trade-off. The increasingly complex models considered allow for more degrees of freedom which, in turn, increase the variance of our absolute benefit estimates. However, as was clear in our simulations, this increased complexity did not always result

in substantial decrease in bias, especially with lower sample sizes and weaker treatment effects. Consequently, in most scenarios the simpler linear interaction model achieved the best bias-variance balance and outperformed the more complex RCS methods, even in the presence of non-linearity in the true underlying relationship between baseline risk and treatment effect. Conversely, the simpler constant treatment effect model was often heavily biased and, despite its lower variance, was outperformed by the other methods in the majority of the considered scenarios.

Increasing the discriminative ability of the risk model reduced RMSE for all methods. Higher discrimination translates in higher variability of predicted risks, which, in turn, allows the considered methods to better capture absolute treatment benefits. As a consequence, better risk discrimination also led to higher discrimination between those with low or high benefit (as reflected in values of c-for-benefit).

The adaptive approach had adequate median performance, following the "true" model in most scenarios. With smaller sample sizes it tended to miss the treatment-baseline risk interaction and selected simpler models (Supplement Section 4). This conservative behavior resulted in increased RMSE variability in these scenarios, especially with true strong linear or non-monotonic deviations. Therefore, with smaller sample sizes the simpler linear interaction model may be a safer choice for predicting absolute benefits, especially in the presence of any suspected treatment-related harms.

A limitation of our simulation study is that we assumed treatment benefit to be a function of baseline risk in the majority of the simulation scenarios, thus ignoring any actual treatment effect modification of individual factors. We attempted to expand our scenarios by considering moderate and strong constant treatment-related harms, applied on the absolute scale, in line with previous work[115]. In a limited set of scenarios with true interactions between treatment assignment and covariates, our conclusions remained unchanged (Supplement, Section 8). Even though the average error rates increased for all the considered methods, due to the miss-specification of the outcome model, the linear interaction model had the lowest error rates. RCS-3 had very comparable performance. The constant treatment effect model was often biased, especially with moderate or strong treatment-related harms. Future simulation studies could explore the effect of more extensive deviations from risk-based treatment effects.

We only focused on risk-based methods, using baseline risk as a reference in a two-stage approach to individualizing benefit predictions. However, there is a plethora of different methods, ranging from treatment effect modeling to tree-based approaches available in more recent literature[56,58,60,64,91,92,109,116].

Many of these methods rely on incorporating treatment-covariate interactions when predicting benefits. An important caveat of such approaches is their sensitivity to overfitting, which may exaggerate the magnitude of predicted benefits. This can be mitigated using methods such as cross-validation or regularization to penalize the effect of treatment-covariate interactions. In the presence of a limited set of true strong treatment-covariate interactions and adequate sample size, treatment effect modeling methods may outperform risk modeling methods. However, often treatment effect modifiers are unknown and the available sample size does not allow for the exploration of a large number of interaction effects. In these cases, risk modeling approaches like the ones presented here can provide individualized benefit predictions that improve on the "one-size-fits-all" overall RCT result. In a previous simulation study, a simpler risk modeling approach was consistently better calibrated for benefit compared to more complex treatment effect modelling approaches[90]. Similarly, when SYNTAX score II, a model developed for identifying patients with complex coronary artery disease that benefit more from percutaneous coronary intervention or from coronary artery bypass grafting was redeveloped using fewer treatment-covariate interactions had better external performance compared to its predecessor[117,118].

Finally, in all our simulation scenarios we assumed all covariates to be statistically independent, the effect of continuous covariates to be linear, and no interaction effects between covariates to be present. This can be viewed as a limitation of our extensive simulation study. However, as all our methods are based on the same fitted risk model, we do not expect these assumptions to significantly influence their relative performance.

In conclusion, the linear interaction approach is a viable option with moderate sample sizes and/or moderately performing risk prediction models, assuming a non-constant relative treatment effect plausible. RCS-3 is a better option with more abundant sample size and when non-monotonic deviations from a constant relative treatment effect and/or substantial treatment-related harms are anticipated. Increasing the complexity of the RCS models by increasing the number of knots does not improve benefit prediction. Using AIC for model selection is attractive with larger sample size.

CHAPTER 4

The EORTC-DeCOG nomogram

Chapter based on *Daniëlle Verver, A. Rekkas, Claus Garbe et al., The EORTC-DeCOG nomogram adequately predicts outcomes of patients with sentinel node–positive melanoma without the need for completion lymph node dissection, European Journal of Cancer, Volume 134, 2020, Pages 9-18, ISSN 0959-8049, https://doi.org/10.1016/j.ejca.2020.04.022*

Abstract

Purpose: Based on recent advances in the management of patients with sentinel node (SN)-positive melanoma, we aimed to develop prediction models for recurrence, distant metastasis (DM) and overall mortality (OM).

Methods: The derivation cohort consisted of 1080 patients with SN-positive melanoma from nine European Organization for Research and Treatment of Cancer (EORTC) centers. Prognostic factors for recurrence, DM and OM were studied with Cox regression analysis. Significant factors were incorporated in the models. Performance was assessed by discrimination (c- index) and calibration in cross-validation across centers. The models were externally validated using a prospective cohort consisting of 705 German patients with SN-positive: 473 trial participants of the German Dermatologic Cooperative Oncology Group study (DeCOG-SLT) and 232 screened patients. A nomogram was developed for graphical presentation.

Results: The final model for recurrence and the calibrated models for DM and OM included ulceration, age, SN tumor burden and Breslow thickness. The models showed reasonable calibration. The c-index for the recurrence, DM and OM model was 0.68, 0.70 and 0.70, respectively, and 0.70, 0.72 and 0.74, respectively, in external validation. The EORTC- DeCOG model identified a robust low-risk group, with all identified low-risk patients (approx- imately 4% of the entire population) having a 5-year recurrence probability of <25% and an overall 5-year recurrence rate of 13%. A model including information on completion lymph node dissection (CLND) showed only marginal improvement in model performance.

Conclusions: The EORTC-DeCOG nomogram provides an adequate prognostic tool for patients with SN-positive melanoma, without the need for CLND. It showed consistent results across validation. The nomogram could be used for patient counselling and might aid in adjuvant therapy decision-making

Introduction

The American Joint Committee on Cancer (AJCC) staging system is the most widely accepted approach to melanoma staging[119,120]. Patients are classified into distinct stages based on the tumor node metastasis criteria where nodal status is based on number of positive lymph nodes after completion lymph node dissection (CLND) in case of a positive sentinel node (SN) or after a therapeutic lymph node dissection in case of clinically apparent nodal disease. Recently there have been many advances in the care of patients with SN-positive melanoma that also affect staging, namely CLND is no longer routine practice as the Multicenter Selective Lymphadenectomy Trial-II (MSLT-II) and the German Dermatologic Cooperative Oncology Group study (DeCOG-SLT) demonstrated no survival benefit for CLND[121–124] and as immune checkpoint inhibition and targeted therapy have been introduced in the adjuvant setting with highly encouraging results[125–128]. Consequently the AJCC staging system is likely to be less appropriate for patients with SN-positive melanoma not undergoing CLND because of decreased discriminatory ability[129] as the number of positive nodes after sentinel lymph node biopsy (SLNB) is not an independent prognostic factor[121,122] (in contrast to involved non-SNs retrieved after CLND[121]). As a result, omitting CLND could result in poorer risk stratification and impaired selection for adjuvant therapy. On the other hand, SN tumor burden has been shown to be an independent predictor of involved non-SNs[130–132], and therefore SN tumor burden may serve as a surrogate.

The objective of the present study was to identify independent prognostic factors in a large European SN- positive melanoma population, using solely information from the primary melanoma and the SLNB, to develop a prediction model for recurrence, distant metastasis (DM) and overall mortality (OM), presented in the form of a nomogram. The resulting model could aid in adjuvant therapy decision-making. The prediction models were externally validated using a large prospective German cohort.

Patients and methods

Cohort characteristics

Derivation cohort

The retrospective derivation cohort consisted of 1080 patients with SN-positive melanoma who underwent SLNB between 1993 and 2008 in one of nine EORTC Melanoma Group centers that have been previously collected and described[129,133–135]. The current study only excluded duplicate cases ($n = 2$), leading to a total of 1078 eligible SN-positive patients. The two

duplicate cases concerned an error in that database. The applied procedures have been described previously[129].

Validation cohort

The prospective German validation cohort involved two sets of patients. The first set consisted of 473 patients who were included in the DeCOG-SLT multicenter randomised phase-3 trial comparing survival between patients with SN-positive melanoma who did or did not undergo CLND[122]. The second set consisted of an additional 219 patients from a single center (University Hospital, Tuebingen) who were initially screened for inclusion in the DeCOG-SLT trial but were not included because of meeting the trial's exclusion criteria (e.g. head and neck melanoma, age >75 years), unwillingness to participate, or no known reason. They also did or did not undergo CLND and were followed and prospectively registered in accordance with similar protocols. All patients had a tumor thickness of at least 1 mm and underwent surgery between 2006 and 2014. The study design, applied procedures and follow-up protocols have been described in detail elsewhere[122]. There was no overlap between the derivation cohort and validation cohort.

Outcomes

Outcomes of interest were first recurrence, first DM and OM. Time to recurrence was calculated from date of SLNB to date of first recurrence or date of death by any cause. Time to first DM was calculated from date of SLNB to date of first DM or date of death by any cause. Time to OM was calculated from date of SLNB to date of death by any cause.

Statistical analysis

The checklist proposed by the AJCC was used for guidance in building a high-quality prediction model[136]. Associations between possible prognostic factors and recurrence were studied with Cox regression analysis. The following eight variables were identified as possible prognostic factors based on clinical experience, literature review and availability of sufficient data: sex, age, ulceration, location, histology, Breslow thickness, total number of SNs removed and total number of positive SNs. To make efficient use of available data, an advanced multiple imputation of missing values strategy (5 imputations) was applied[137]. This was done separately for each derivation center to avoid using information of missingness in cross-validation. The possible non-linearity of continuous variables was modeled by logarithmic transformation. Independent prognostic factors were selected with multivariable backwards selection. Linear predictor values (the sum of truncated predictor values times their predictor effects) were scaled and rounded to a risk score with integer values between 0 and 100. Because recurrence, DM and OM are strongly related, the final recurrence

prediction model based on data from all nine EORTC centers was used as a basis for predicting DM and OM, where the baseline hazard and the slope of the recurrence prediction model were calibrated to DM and OM[138]. This approach is beneficial as it provides a unique risk score for each individual that translates into probabilities of all outcomes of interest, instead of developing three independent prediction models. To test the validity of our approach, we did develop these independent models and compared them with the calibrated models. The absolute risk prediction of each outcome was plotted against the risk score. To reduce overestimation of events occurring in patients with extremely high scores, scores were truncated at an integer of 23, corresponding to the 99th percentile of score distribution. Model performance was assessed by examining discrimination and calibration. Discrimination was measured using the concordance index (c-index); the closer to 1, the better the discrimination, and a value of 0.5 indicates that the model is no better than a chance[139]. Calibration was assessed visually by plotting the predicted probability against the actual observed frequency in quintiles of predicted outcomes. A 45 line indicates perfect calibration (when the predictive value of the model perfectly matches the patient's actual risk). Any deviation above or below the 45 line indicates under-prediction or over-prediction, respectively. A nomogram was developed for graphical presentation of the models. To evaluate generalizability of the models across different centers, an internal-external cross-validation was performed in which the model was fitted using data from eight centers and validated in the center that was left out[140]. In addition we performed external validation using the prospective German cohort. We first needed to develop a model for recurrence where we replaced the continuous variable SN tumor burden with the categorical substitute used in the prospective German cohort (single cells, <0.5 mm, 0.5e1.0 mm, >1.0e2.0 mm, >2.0e5.0 mm and >5.0 mm). For the derivation cohort, single cells were defined as <0.1 mm according to the Rotterdam criteria[141]. Single cells in the validation cohort were not specifically defined, but as the Rotterdam criteria were used for measuring SN tumor burden, definitions are likely to correlate. The performance of this altered model was compared with the final recurrence model used for the nomogram. Subsequently the altered model was externally validated with the 692 patients from the prospective German cohort. To test how much the information on additional positive nodes retrieved after CLND would add to the discrimination of the prediction model, we also developed a prediction model in which the variable, additional positive nodes after CLND, was added. This model was based on 1015 patients that underwent CLND in the derivation cohort.

Furthermore we calculated the model performance for recurrence, DM and

OM of the AJCC 7th edition classification, AJCC 8th edition classification and the simple classification that was published previously (i.e. absent/present ulceration and low/high SN tumor burden) was tested[129]. Lastly the observed outcomes per group for all classifications were estimated using the Kaplan Meier analysis. All statistical tests were two-sided, with a $P < 0.05$ considered statistically significant. All statistical analyses were performed using SPSS version 22.0 (IBM, Armonk, New York, USA) and R (version 2.15, R Foundation for Statistical Computing, Vienna, Austria, 2011).

Results

The retrospective derivation cohort consisted of 1078 and the prospective validation cohort of 692 patients with SN-positive. Patients in the validation cohort had less extensive disease in terms of Breslow thickness, number of positive SNs and tumor burden in the SN compared with those in the derivation cohort (Table 1).

In the derivation cohort, recurrence at five years occurred in 496 patients (46.0%), DM in 437 patients (40.5%) and OM in 364 patients (33.8%). Median follow-up time for all survivors was 106 months (interquartile range [IQR] 61e130 months). In the prospective validation cohort, recurrence at five years occurred in 267 patients (38.6%), DM in 223 patients (32.2%) and OM in 174 patients (25.1%). Median follow-up time for all survivors was 66 months (IQR: 48-94 months).

Models for recurrence, distant metastasis and overall mortality

The final multivariable Cox model for recurrence after backwards selection included four independent prognostic factors: ulceration, age, Breslow thickness and SN tumor burden (Table 2). Logarithmic transformation of the continuous variables adequately represented their effects. The c-index for the final recurrence model was 0.68 (95% confidence interval [CI]: 0.65e0.70). In cross-validation, the recurrence model was reasonably calibrated across nine center in general, only in smaller centers there was substantial underestimation of the risk (Fig. S1).

The association between linear predictors of recurrence and DM was of the same size (calibration slope: 1.01, 95% CI: 0.87-1.16). The c-index for the calibrated model for DM was 0.70 (95% CI: 0.67-0.72) and was reasonably calibrated across nine in cross-validation (Fig. S2). The performance of this calibrated model, based on the baseline hazard and the slope of the recurrence model, was similar to that of the independently developed prediction model for DM (c-index: 0.70, 95% CI: 0.68-0.73)

The association between linear predictors of recurrence and OM was of the same size (calibration slope: 1.04, 95% CI: 0.88-1.20). The c-index for the calibrated model for OM was 0.70 (95% CI: 0.67-0.73), and was reasonably calibrated across nine centers in cross-validation (Fig. S3). The performance of this calibrated model was similar to that of the independently developed prediction model for OM (c-index: 0.70, 95% CI: 0.68-0.73).

A four-item risk score was developed, assigning points to each prognostic factor based on the magnitude of association with recurrence. A nomogram to calculate the score and the risk of recurrence, DM and OM is presented in Fig. 1. The scores were divided into four risk groups based on the 5-year probability of recurrence: <25% (low risk; score 6-9; 4.1% of the population); 25-50% (intermediate risk; score 10e15; 52.9% of the population); 50-75% (high risk; score 16-19; 33.2% of the population); and >75% (very high risk; score 20e23; 10.0% of the population). The observed outcomes for recurrence, DM and OM per risk group are shown in Table 3.

External validation

For external validation purposes, an altered recurrence model was developed using the categorized SN tumor burden variable used in the prospective German cohort (Table S1). This altered model showed similar performance compared with the final recurrence model (c- index 0.68, 95% CI: 0.65-0.70). In external validation, the c-index for the altered recurrence model was 0.70 (95% CI:tumor 0.67-0.74), for DM 0.72 (95% CI: 0.68-0.75) and for OM 0.74 (95% CI: 0.71-0.78). The calibration plots indicate good calibration, though there may be slight underestimation for higher-risk patients in the recurrence and OM models (Fig. S4).

Additional prognostic value of CLND

An extended model for recurrence was created by adding the variable, number of additional positive nodes after CLND, to the final recurrence model. This extended model for recurrence had a c-index of 0.69 (95% CI: 0.67e0.72). The calibrated extended models for DM and OM showed c-indices of 0.72 (95% CI: 0.69e0.74) and 0.72 (95% CI: 0.69e0.75), respectively.

Simple classification

A simplified version of the model stratifies patients into four groups based on ulceration and SN tumor burden: 1) absent ulceration and $\leq$ 1.0 mm; 2) absent ulceration and >1.0 mm; 3) present ulceration and $\leq$ 1.0 mm and 4) present ulceration and >1.0 mm. The c-indices for this classification in predicting recurrence, DM and OM were 0.63 (95% CI: 0.61e0.65), 0.64 (95% CI: 0.62e0.67) and 0.64 (95% CI: 0.61e0.67), respectively. The observed outcomes

for recurrence, DM and OM per risk group are shown in Table 3.

The American Joint Committee on Cancer (AJCC) classifications

Patients were classified based on the 7th AJCC classification into IIIA $\leq$ 1.0 mm, IIIA >1.0 mm, IIIB and IIIC and based on the 8th edition into IIIA $\leq$ 1.0 mm, IIIA >1.0 mm, IIIB, IIIC and IIID. The c-indices for predicting recurrence, DM and OM for the 7th AJCC edition were 0.61 (95% CI: 0.59-0.63), 0.62 (95% CI: 0.60-0.65) and 0.62 (95% CI: 0.59-0.65), respectively, and for the 8th AJCC edition 0.62 (95% CI: 0.59-0.64), 0.63 (95% CI: 0.60-0.65) and 0.63 (95% CI: 0.61-0.66), respectively. The observed outcomes for recurrence, DM and OM for both AJCC classifications are shown in Table 3. A cross-table comparing the patients staged in accordance with the AJCC classifications and the risk groups based on the EORTC-DeCOG model is illustrated in Table 4. An overview of c-indices for all the different models is presented in Table 5.

Discussion

The present study developed and validated a nomogram to predict five-year recurrence, DM and OM in patients with SN-positive melanoma, by solely using information from the primary melanoma and SLNB. The resulting patient-specific probabilities could be used to tailor adjuvant therapeutic strategies for patients with SN-positive melanoma, without the prerequisite to undergo CLND and thereby avoiding potential significant morbidity. The greatest contemporary value of our prognostic nomogram is the possibility of identifying patients at sufficiently low risk for recurrence, DM and OM in whom adjuvant therapy could be omitted.

Although the FDA and EMA pragmatically approved adjuvant therapy for all stage-III patients, it is still under debate which patients should not be considered candidates. Patients with stage IIIA $\leq$ 1.0 mm (AJCC 7th edition) were considered low risk in most adjuvant therapy trials and were therefore not included (one even excluded all IIIA patients)[125–127,142,143]. The current study indicates that when the AJCC 8th edition criteria are used for defining IIIA $\leq$ 1.0 mm instead of the 7th edition, it results in improved selection of low- risk patients in terms of predicted prognosis (e.g. 5-year recurrence probability of 27% versus 32%, respectively). A recent study also showed that including SN tumor burden to the 8th AJCC staging system has crucial prognostic relevance[144]. Of note our EORTC-DeCOG model is able to identify an even more robust low-risk group, as all identified low-risk patients (which approximately concerned 4% of the entire population after imputation) had a 5-year recurrence probability of <25% and an overall 5-year observed recur-

rence rate of 13%. However, identifying more robust low-risk groups comes at the cost of fewer patients being assigned low risk (see Table 4). Nonetheless a major advantage of our EORTC-DeCOG model is that it provides a more continuous type of predicted probabilities. As a result it is possible to derive risk groups based on outcome probabilities and/or risk scores (e.g. low risk; scores 6–9; recurrence probability of <25%) which is in contrast to the AJCC classifications where exact patient/ tumor characteristics define the risk groups (e.g. IIIA $\leq$ 1.0 mm: T1a/b-T2a + N1a-N2a with $\leq$ 1.0 mm SN tumor burden). In the current study we choose to derive risk groups based on the recurrence probability, as this seems the most relevant outcome in the context of selecting patients for adjuvant therapy; other cut-off values and/or outcomes are possible. In conclusion, the EORTC-DeCOG model not only outperforms the AJCC classifications in terms of overall model discrim- ination (see Table 5), but also seems to be able to identify a more robust low-risk group in whom it may be justified to forego adjuvant therapy.

The previously published simplified model, based on ulceration and SN tumor burden, harbored the least performance, though still reasonable, and showed similar predicted prognosis for the low-risk group as the 7th AJCC edition. Whether to implement a more complex model versus a less robust model is a balance between performance and simplicity. In our opinion, the simple model could serve as an easy user-friendly prognostic tool for daily clinical practice and to generally inform patients, but for more adequate risk estimates and decisions upon (adjuvant) treatment, we advocate using the comprehensive EORTC-DeCOG model. Noteworthy, besides the common prognostic factors (i.e. ulceration, Breslow thickness and SN tumor burden), the current study also identified increasing age as an independent prognostic factor for recurrence, DM and OM. This finding is supported by other studies reporting on the significance of the patient's age[145].

Stratifying for ulceration and SN tumour burden only was previously demonstrated to yield similar discriminatory ability for melanoma-specific mortality as stratifying for AJCC sub-stages which included information on nodal status after CLND[129]. The additional value of non-SN status retrieved after CLND was also tested in the current study, by developing an extended model. This model showed only marginal improvement in performance (e.g. c-index for the recurrence model increased from 0.68 to 0.69), thereby indicating that omitting CLND has very limited consequences for prognostication if SN tumor burden is taken into account.

This study has several limitations. First is the retrospective design of the derivation cohort, which has inherent biases. However, the models proved to be successful in external validation. Performance was comparable between the

derivation and prospective validation cohort, even though the latter cohort included patients with relatively better prognosis (e.g. less extensive disease) and largely represents a clinical trial population. Adjuvant interferon-α therapy was intended in approximately 60% of the patients included in the DeCOG-SLT trial, which is another possible limitation[122]. It could have potentially influenced outcomes, especially in patients with ulcerated melanomas as ulceration seems to be a predictive factor for IFN sensitivity[146,147]. Furthermore, it is unknown how many patients in the validation cohort received effective novel therapy after recurrence. Because patients were included from 2006 through 2014, it is likely some patients did. As patients in the derivation cohort were included from 1993 through 2008, novel therapies probably had limited effect. To date, no novel biomarker has been validated that suffices to predict long-term clinical benefits and subsequently could be incorporated in the models, despite efforts in this drection (e.g. PD-L1)[148]. In addition, other prognostic factors such as mitotic rate or microsatellites could not be incorporated in the present models because of insufficient data. Another limitation is the inadequate representation of patients with SN-positive with a head and neck melanoma in both cohorts. For the validation cohort this is largely explained as it was an exclusion criterion in the DeCOG-SLT trial, and for the derivation cohort this might be partially explained by the historical concerns of poor safety, accuracy and prognostication. Similar numbers (~5%) have been reported in other European cohorts[149,150], while particularly American cohorts have reported higher numbers (>10%)[121,151]. With the introduction of adjuvant therapies, the number of performed SLNBs in head and neck melanomas is likely to increase.

Considering the advances in the management of patients with SN-positive melanoma, it becomes highly relevant to have a prediction model that provides precise patient-specific probabilities based on solely factors from the primary melanoma and the SLNB. The EORTC-DeCOG nomogram is the first that meets these demands, and as a result it could be used for patient counselling and assist in trial design. In addition it might aid in adjuvant therapy decision-making. To facilitate its use, an online calculator has been developed and can be accessed at https://www. evidencio.com/models/show/2010.

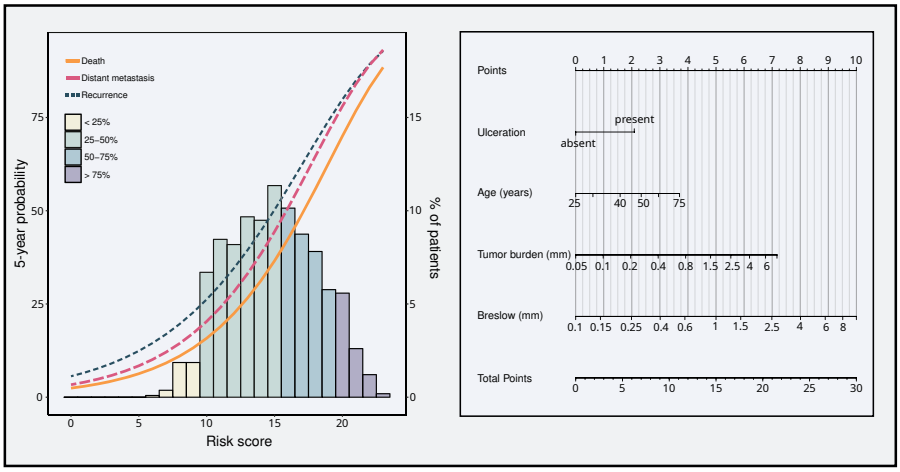

Figure 4.1: Nomogram and risk distribution. The curves refer to predicted recurrence, distant metastasis or overall mortality at 5 years. The histogram refers to the risk score distribution in the cohort; each bar represents the proportion of patients in the cohort that was assigned that specific score. The histogram was divided in four risk groups based on the risk of recurrence: low risk: <25%, intermediate risk: 25e50%, high risk: 50e75% and very high risk: >75%. The nomogram incorporates four factors: ulceration, age, SN tumor burden and Breslow thickness. To calculate an individual's probability of 5-year recurrence, distant metastasis and overall mortality, values for the prognostic factors must be determined first (for example: absent ulceration, 35 years, SN tumor burden 0.8 mm and Breslow thickness 1.0 mm). Second, for each value the corresponding points can be obtained by drawing a line from each value towards the point axis (in example: 0, 1, 4 and 5 points, respectively). Third, the points must be added up to obtain the total risk score (in example: risk score of 10). Finally, the 5-year recurrence, distant metastasis and overall mortality probability can be read by moving vertically from the x-axis (total risk score) to the predicted risk curves and corresponding probabilities on the left y-axis (in example: 26% for recurrence, 20% for distant metastasis and 16% for overall mortality). The percentage of patients in the entire population (1078) that also had a total risk score of 10 can be determined from the histogram, as well as the corresponding percentage of patients on the right y-axis (in example: 7%).

CHAPTER 5

COVID outcome prediction in the emergency department (COPE)

Chapter based on *van Klaveren D, Rekkas A, Alsma J, et alCOVID outcome prediction in the emergency department (COPE): using retrospective Dutch hospital data to develop simple and valid models for predicting mortality and need for intensive care unit admission in patients who present at the emergency department with suspected COVID-19BMJ Open 2021;11:e051468. doi: 10.1136/bmjopen-2021-051468*

Abstract

Objectives: Develop simple and valid models for predicting mortality and need for intensive care unit (ICU) admission in patients who present at the emergency department (ED) with suspected COVID-19.

Design: Retrospective.

Setting: Secondary care in four large Dutch hospitals.

Participants: Patients who presented at the ED and were admitted to hospital with suspected COVID-19. We used 5831 first-wave patients who presented between March and August 2020 for model development and 3252 second-wave patients who presented between September and December 2020 for model validation. Outcome measures We developed separate logistic regression models for in-hospital death and for need for ICU admission, both within 28 days after hospital admission. Based on prior literature, we considered quickly and objectively obtainable patient characteristics, vital parameters and blood test values as predictors. We assessed model performance by the area under the receiver operating characteristic curve (AUC) and by calibration plots.

Results: Of 5831 first-wave patients, 629 (10.8%) died within 28 days after admission. ICU admission was fully recorded for 2633 first-wave patients in 2 hospitals, with 214 (8.1%) ICU admissions within 28 days. A simple model with age, respiratory rate, C reactive protein, lactate dehydrogenase, albumin and urea captured most of the ability to predict death. COPE was well calibrated and showed good discrimination for mortality in second-wave patients (AUC in four hospitals: 0.82 (95% CI 0.78 to 0.86); 0.82 (95% CI 0.74 to 0.90); 0.79 (95% CI 0.70 to 0.88); 0.83 (95% CI 0.79 to 0.86)). COPE was also able to identify patients at high risk of needing ICU admission in second-wave patients (AUC in two hospitals: 0.84 (95% CI 0.78 to 0.90); 0.81 (95% CI 0.66 to 0.95)).

Conclusions: COPE is a simple tool that is well able to predict mortality and need for ICU admission in patients who present to the ED with suspected COVID-19 and may help patients and doctors in decision making.

Background

The COVID-19 pandemic is putting extraordinary pressure on emergency departments (EDs), clinical wards and intensive care units (ICUs). Clinical prediction models for COVID-19 outcomes have the potential to support decision making about hospital admission. Existing models that predict mortality for non-trauma patients presenting to the ED are unlikely to be well calibrated and optimally discriminating for patients with COVID19[152]. Most currently available models specifically developed for patients with COVID-19 that were assessed with the prediction model risk of bias assessment tool contain a high risk of bias[153–155]. The most common reasons were non-representative selection of control patients, exclusion of patients in whom the event of interest was not observed by the end of the study, high risk of model overfitting and vague reporting. Additionally, the description of the study population or intended use of the models was often missing, and calibration of the model predictions was rarely assessed.

The recently proposed 4C Mortality Score is probably at low risk of bias, but was derived from a selected population of patients admitted to UK hospitals who were seriously ill (mortality rate of 32.2%). Predictors included the number of comorbidities and the Glasgow Coma Scale, items that are not easily and unambiguously obtained for patients with suspected COVID-19 at EDs everywhere[156,157]. Similarly, the promising risk scores Veterans Health Administration COVID-19 (VACO) and COVID-GRAM—predicting 30-day mortality in positively tested patients and critical illness in hospitalised patients, respectively—require knowledge on pre-existing comorbidities[158,159]. The COVID-GRAM model also requires chest radiography results.

We aimed to develop and validate a simple and valid model for predicting mortality and the need for ICU in all patients who are suspected to have COVID-19 when presenting at the ED. To facilitate implementation in clinical practice, we only included quickly and objectively obtainable patient characteristics, vital parameters and blood test values.

Methods

Population

Nineteen large Dutch hospitals were requested to supply anonymised retrospective data on the cohorts of patients with COVID-19 who were admitted to their hospital. Of those hospitals, Catharina Hospital Eindhoven, Zuyderland Medical Center Heerlen, Isala Clinics Zwolle, Erasmus University Medical Center Rotterdam and Antonius Hospital Sneek supplied these data. The data from Antonius Hospital Sneek were not used in the analyses, because of large

proportions of missing predictor values.

For model development, we used the data of patients who presented at the ED and were admitted to the hospital with suspected COVID-19 in the first wave of the pandemic, that is, from March up to and including August 2020. Patients being transferred to other hospitals were excluded since information on outcomes was missing. For model validation, we used data of patients who presented at the ED and were admitted to the hospital with suspected COVID-19 in the second wave of the pandemic, that is, from September up to and including December 2020. Potential multiple hospital admissions of the same patient were considered as independent hospital admissions.

Outcomes

The outcomes of interest were: (1) in-hospital death or transfer to a hospice within 28 days after hospital admission and (2) admission to ICU within 28 days after hospital admission.

Predictors

Based on prior literature, we included patient characteristics (sex, age, body mass index), vital parameters (oxygen saturation, systolic blood pressure, heart rate (HR), respiratory rate (RR), body temperature) and blood test values (C reactive protein (CRP), lactic dehydrogenase (LDH), D-Dimer, leucocytes, lymphocytes, monocytes, neutrophils, eosinophils, Mean Corpuscular Volume (MCV), albumin, bicarbonate, sodium, creatinine, urea), all measured at ED admission, as potential predictors[153]. Furthermore, we included the month of admission to capture potential changes in outcomes over time. In case of multiple measurements for the same patient, we used the first measurement after presentation at the ED. We used multivariate imputation by chained equations (R-packages mice) for multiple imputation of missing predictor values[137,160]. Multiple imputation in the validation data was undertaken separately from multiple imputation in the development data to ensure fully independent model validation.

Model development

Logistic regression was used to analyse associations between predictors and outcomes. We decided on including non-linear transformations of potential predictors on the basis of a full model with a restricted cubic spline (three knots; two regression coefficients) for each continuous predictor[161,162]. Based on Wald statistics, we selected the most promising predictors into a parsimonious model for easy use in clinical practice. To prevent overfitting, we used bootstrap validation—including the same variable selection strategy to mimic our modelling strategy—to estimate a uniform shrinkage factor[162]. The regres-

sion coefficients of the final model were multiplied by this shrinkage factor, and the model intercept was adjusted to ensure overall calibration of the model. We used the R-package rms (Regression Modelling Strategies) for regression analyses[160,163].

Model validation

Model performance was assessed with temporal validation in second wave patients, in each of the four separate hospitals. We assessed discriminative ability with the area under the receiver operating characteristic curve (AUC) and calibration with calibration plots of five equally sized groups of predicted risk, calibration intercepts and calibration slopes. The model-based concordance (mbc) was used to understand the impact of potential differences in casemix heterogeneity between the development and validation data on discriminative ability[164].

Patient and public involvement

Patients were not directly involved in the design of this study. The outcome of interest and the potential predictors were selected up front by a group of hospital physicians caring for patients with COVID-19 (ED, internal medicine, pulmonary medicine, ICU). Since we retrospectively collected data, patients were not burdened by our study. In future research, we will convene multi-stakeholder panels of approximately 12 members including patients with COVID-19, relatives, hospitals physicians caring for patients with COVID-19, palliative care physicians and ethicists, with the aim to develop a full understanding of how the models may best support patients and clinicians in making critical patient-centred decisions.

Results

Population and outcomes

The database contained 5912 patients who presented at the ED from March up to and including August 2020 and who were admitted to the hospital with a suspicion of COVID-19. Of those patients 81 (1.4%) were excluded because of a transfer to other hospitals (outcome not recorded). The development data included 5831 patients of whom 629 (10.8%) died, 5070 (86.9%) were discharged within 28 days after hospital admission, and 132 (2.3%) were still in hospital at 28 days after admission. Patients who died—in comparison with patients who were discharged—tended to be more often male (64% vs 56%), at older age (median 78 vs 69), with higher RR (median 23 vs 19) and HR (median 93 vs 90), lower oxygen saturation (median 94.1 vs 96.0), higher blood levels of CRP (median 91 vs 43), LDH (median 338 vs 237), creatinine (median

102 vs 82) and urea (median 9.6 vs 6.2) and lower blood levels of lymphocytes (median 0.80 vs 1.10) and albumin (median 36 vs 40) (Table 5.1). Similar patterns were seen in 3252 patients who were admitted to hospital in the second wave of the pandemic from September up to and including December 2020, of whom 326 (10.0%) died, 2854 (87.8%) were discharged within 28 days after admission, and 72 (2.2%) were still in hospital at 28 days after admission.

Admission to ICU was fully recorded—including ICU admissions at a later time point than the initial hospital admission—for 2633 patients in 2 hospitals (214 ICU admissions within 28 days (8.1%)) in the first wave of the pandemic. Patients who were admitted to the ICU—in comparison with patients who were discharged or died without being admitted to the ICU—tended to be more often male (68% vs 57%), with higher RR (median 23 vs 19) and HR (median 91 vs 88), lower oxygen saturation (median 95.0 vs 95.8), higher blood levels of CRP (median 88 vs 47), LDH (median 318 vs 234), creatinine (median 93 vs 84) and urea (median 7.1 vs 6.6) and lower blood levels of albumin (median 38 vs 40) (Table 5.2). In contrast with patients who died, patients who were admitted to the ICU were not older than patients who were discharged (median 68 vs 71), probably due to decisions not to admit frail patients to the ICU. Patterns were similar in 1466 patients (86 ICU admissions within 28 days (5.9%)) who were admitted to these 2 hospitals in the second wave of the pandemic.

Prediction of death

Patients who were admitted in the first month of the pandemic in the Netherlands, that is, in March 2020, were at substantially increased risk of death (Table 5.3: multivariable OR 1.99; 95% CI 1.61 to 2.47). All models included this correction factor for the first month, to avoid overestimation of risk after the first month of the pandemic. Consequently, to avoid overestimation of the discriminative ability, we limited validation of models in the development data to patients who were admitted from April 2020 onward.

D-dimer concentration in the blood, measured to detect thrombosis, was not analysed in the regression analysis, because 64% and 76% were missing in the development and validation data, respectively (Table 5.1). Based on a full model with restricted cubic splines of all potential variables, we decided to transform all biomarkers and RR with the natural logarithm, while keeping all other predictor effects linear. Some strong univariable associations with death—for example of logarithmically transformed lymphocytes and creatinine (Table 5.3; Wald statistics 48 and 133, respectively)—were very weak in multivariable analysis (Table 5.3; Wald statistics 0 and 4, respectively). The predictive ability of the resulting full multivariable regression model was

mainly driven by age, LDH, urea, RR, CRP, Albumin, oxygen saturation and bicarbonate (ORs and Wald statistics in Table 5.3). A simple model—named COVID outcome prediction in the emergency department (COPE)—with linear age and logarithmic transforms of RR, CRP, LDH, albumin and urea captured most of the ability to predict death within 28 days (Table 5.3; Figure 5.1). Based on internal bootstrap validation, we applied a shrinkage factor of 0.93 to the regression coefficients.

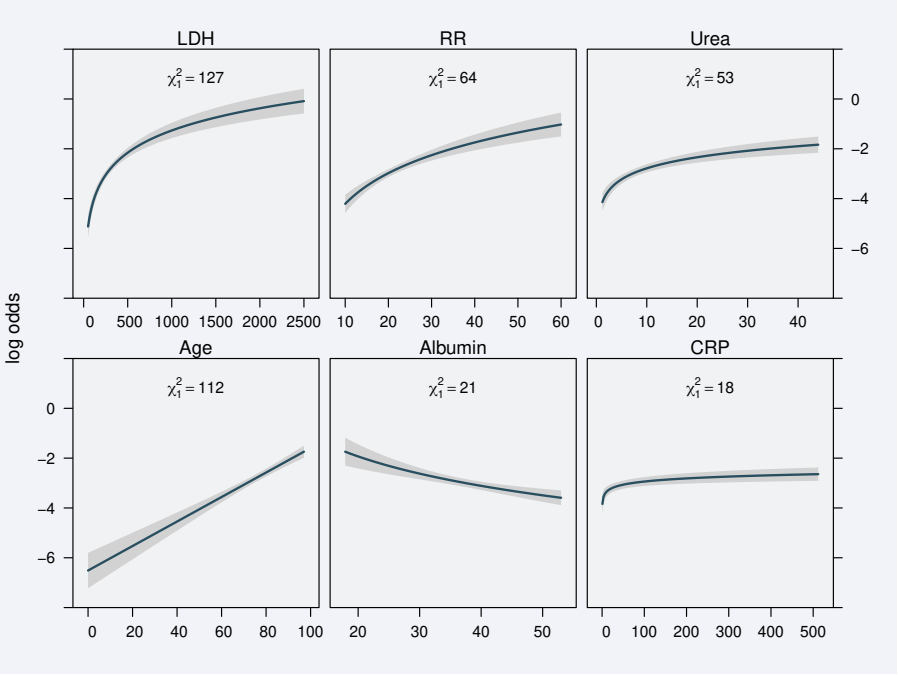

Figure 5.1: Multivariable effects of continuous predictors of death within 28 days predictions of the logarithm of the odds by continuous predictor levels, with other predictor levels set to the median. Wald statistics are listed within each plot to express variable importance (higher is better). CRP, C reactive protein; LDH, lactic dehydrogenase; RR, respiratory rate.

COPE showed good discrimination for predicting death in 4498 patients who were admitted from April up to and including August 2020 in the first wave (online supplemental Figure 1); AUC in 4 hospitals 0.85 (95% CI 0.81 to 0.88)); 0.81 (95% CI 0.71 to 0.91); 0.86 (95% CI 0.82 to 0.90); 0.85 (95% CI 0.81 to 0.88)) and, more importantly, in the validation sample of 3235 patients who were admitted in the second wave from September up to and including December 2020 (Figure 5.2; AUC in four hospitals: 0.82 (95% CI 0.78 to 0.86);

0.82 (95% CI 0.74 to 0.90); 0.79 (95% CI 0.70 to 0.88); 0.83 (95% CI 0.79 to 0.86)). The decrease in AUC over time was partly driven by less case mix heterogeneity—expressed by a lower model-based AUC (mbc)—of second wave patients (Figure 5.2; mbc in four hospitals: 0.81; 0.82; 0.81; 0.82) as compared with first wave patients (online supplemental Figure 1); mbc in four hospitals 0.82; 0.85, 0.83, 0.84). COPE was well calibrated in second wave patients of each of the four hospitals, both on average—expressed by hospital-specific calibration intercepts: 0.08 (95% CI −0.15 to 0.30); −0.17 (95% CI −0.65 to 0.30); −0.01 (95% CI −0.40 to 0.39); −0.12 (95% CI −0.30 to 0.07)—and by predicted risk levels—expressed by hospitalspecific calibration slopes: 1.09 (95% CI 0.86 to 1.31); 0.90 (95% CI 0.49 to 1.32); 0.91 (95% CI 0.57 to 1.25); 0.97 (95% CI 0.79 to 1.14) (Figure 5.2).

When stratifying second wave patients according to a mortality risk threshold equal to the event rate (10%), COPE assigned high risk to 246/326 patients who actually died (76% sensitivity, ie, 24% false negatives) and low risk to 2086/2926 patients who actually survived (71% specificity, ie, 29% false positives). With a 5% risk threshold, the sensitivity increased to 93% while the specificity decreased to 49%. Based on a 20% risk threshold, the sensitivity decreased to 49% while the specificity increased to 89%.

Prediction of need for ICU admission

The probability of being admitted to the ICU was decreasing with age after the age of 70 (OR of age 80 vs 70–79: 0.17 (95% CI 0.10 to 0.30)), likely reflecting the decision not to admit older patients to the ICU. When adjusting for this decreasing age effect after the age of 70—by including a linear spline with a knot at age 70 in the regression model (online supplemental Figure 2)—the strongest predictors of death were also predictive of ICU admission within 28 days, but associations were generally weaker for the latter (table 4 vs table 3). In patients below the age of 70, admitted from April up to and including August 2020, a model with the linear predictor of death calibrated to ICU admission had similar discriminative ability to a model that refitted all the predictor effects (AUC 0.71 for both models). For robustness, we implemented the calibrated model, also adjusted for a linearly decreasing age effect after the age of 70, and not the refitted model (calibration slope 0.60; 95% CI 0.49 to 0.70) into COPE for predicting ICU admission. To predict the need for ICU admission of future patients over the age of 70 COPE ignores the decreasing age effect after the age of 70, since the observed ICU admission rate is probably an inaccurate estimate of the medical need for ICU admission. By fitting a linearly decreasing age effect in patients over the age of 70 which is not applied when predicting for future patients, predictions of ICU admission after the age of 70 are based on an extrapolation of the observed age effect on

ICU admission in patients below the age of 70. Due to the weaker predictor effects, the discriminative ability of COPE was more moderate for predicting ICU admission than for predicting death (online supplemental Figure 3); AUC in two hospitals: 0.66 (95% CI 0.58 to 0.74); 0.79 (95% CI 0.69 to 0.88)). Although COPE significantly overestimated ICU admission in second wave patients (Figure 5.3; calibration intercept in two hospitals: −0.41 (95% CI −0.77 to −0.05); −0.72 (95% CI −1.34 to −0.11)), it was better able to identify the patients at high risk of needing ICU admission, as expressed by higher discriminative ability (Figure 5.3; AUC in two hospitals: 0.84 (95% CI 0.78 to 0.90); 0.81 (95% CI 0.66 to 0.95)) and substantially stronger predictor effects (calibration slope in two hospitals: 1.55 (95% CI 1.03 to 2.06); 1.53 (95% CI 0.60 to 2.46)).

Model presentation

The resulting COPE models for predicting death as well as need for ICU admission within 28 days after hospital admission (formulas in Table 5.5) are implemented as a publicly accessible web-based application (https://mdmerasmusmc. shinyapps.io/COPE/) and as independent mobile apps ('COPE Decision Support'). For optimal transparency, the web and mobile applications include a detailed description of the derivation of COPE (online supplemental file 1), descriptions of the data that were used for development and validation of COPE, and calibration plots of temporal validation in the separate hospitals. According to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis checklist (online supplemental table 1), all relevant items are covered in this manuscript, except for the availability of data sets. (***REFS: 15 16***) The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to data transfer agreements with each of the contributing hospitals.

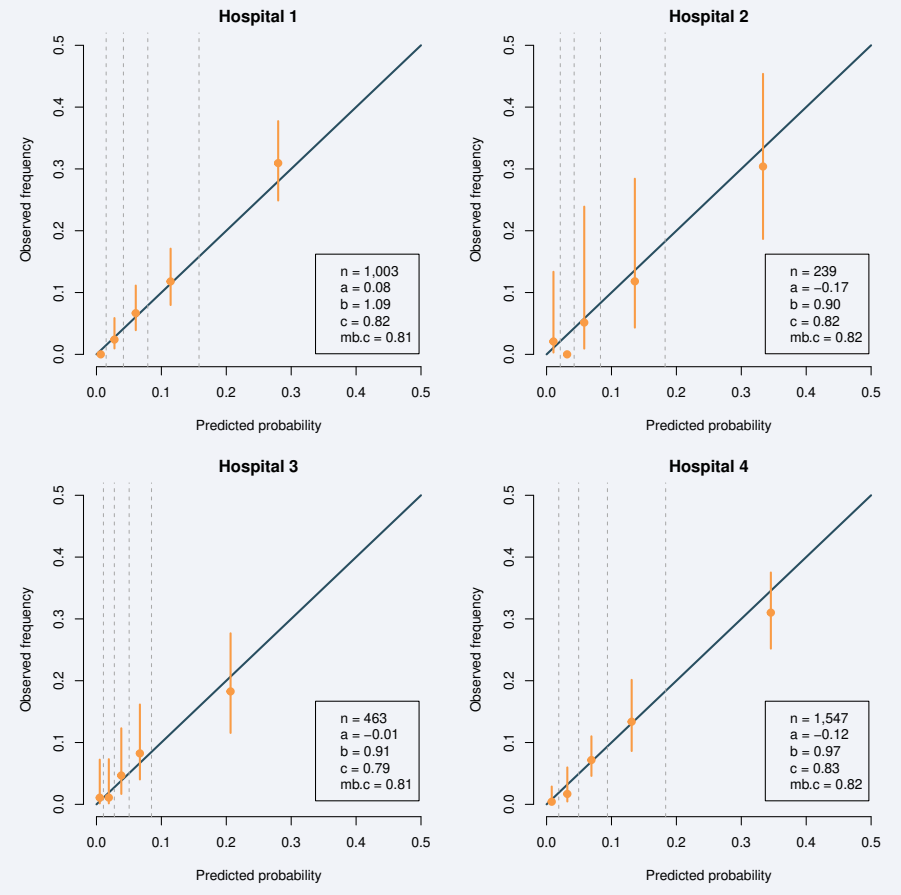

Figure 5.2: Temporal validation: Performance of COPE for predicting death in second wave patients calibration plots of patients who were admitted since September 2020 in four separate Dutch hospitals. n is number of patients; a=calibration intercept (0 is perfect); b=calibration slope (1 is perfect); c=AUC (0.5 is useless; 1 is perfect); mbc=model-based AUC. AUC, area under the curve; COPE, COVID-19 outcome prediction in the emergency department.

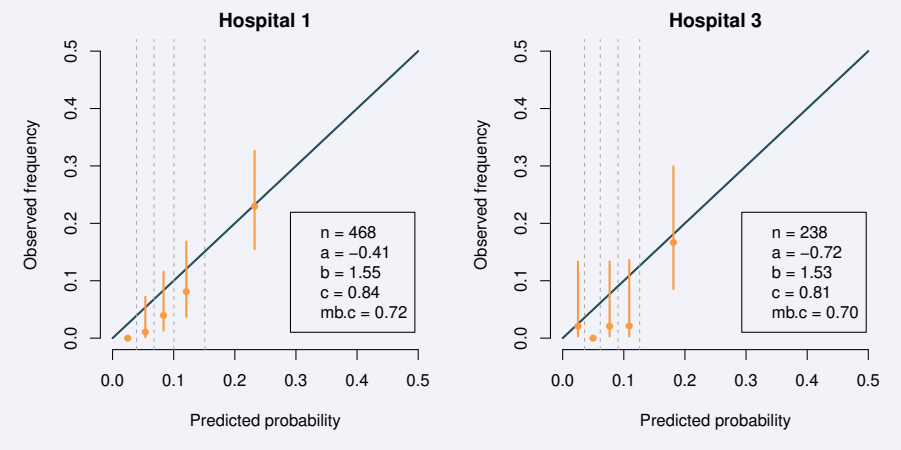

Figure 5.3: Temporal validation: performance of COPE for predicting ICU admission in second wave patients calibration plots of patients who were admitted since September 2020 in two separate Dutch hospitals. n is number of patients; a=calibration intercept (0 is perfect); b=calibration slope (1 is perfect); c=AUC (0.5 is useless; 1 is perfect); mbc=model-based AUC. AUC, area under the curve; COPE, COVID-19 outcome prediction in the emergency department; ICU, intensive care unit.

	% missing	All patients	Discharged or dead	In hospital	ICU admission
A. Development data		n = 5831	n = 5070	n = 132	n = 629
Male sex	0	57%	56%	64%	64%
Age (years)	0	70 (58, 80)	69 (56, 78)	71 (62, 79)	78 (70, 84)
BMI (kg/m²)	58	26 (23, 30)	26 (23, 30)	25 (23, 29)	26 (23, 30)
HR (bpm)	39	90 (78, 103)	90 (78, 103)	87 (76, 99)	93 (80, 107)
SBP (mmHg)	42	133 (118, 150)	133 (119, 151)	136 (115, 152)	131 (114, 145)
RR (/min)	42	19 (16, 23)	19 (16, 23)	20 (17, 24)	23 (19, 28)
Saturation (%)	41	95.8 (94, 97.5)	96 (94.3, 97.8)	95.4 (93.6, 97)	94.1 (91.9, 96)
Temperature (°C)	40	37.3 (36.7, 38.1)	37.3 (36.7, 38)	37.5 (37, 38.1)	37.4 (36.7, 38.1)
CRP (mg/L)	7	48 (10, 118)	43 (8, 110)	85 (34, 160)	91 (41, 180)
D dimer (µg/L)	64	1100 (527, 2545)	1028 (504, 2300)	1950 (719, 8110)	2100 (949, 4772)
LDH (U/L)	18	244 (200, 322)	237 (197, 302)	300 (234, 422)	338 (253, 492)
Leucocytes (x10^9/L)	7	9.1 (6.7, 12.7)	9.1 (6.7, 12.6)	9.4 (6.5, 12.5)	9.2 (6.4, 14)
Lymphocytes (x10^9/L)	16	1.04 (0.66, 1.6)	1.1 (0.7, 1.7)	0.8 (0.54, 1.3)	0.8 (0.51, 1.3)
Albumin (g/L)	15	39 (36, 42)	40 (36, 43)	37 (34, 41)	36 (33, 39)
Bicarbonate (mmol/L)	45	23.6 (21, 26)	24 (22, 26)	23 (21, 26)	22 (20, 25)
Creatinine (µmol/L)	8	84 (66, 111)	82 (65, 107)	89 (66, 111)	102 (75, 153)
Eosinophils (x10^9/L)	26	0.03 (0, 0.1)	0.03 (0.01, 0.1)	0.03 (0, 0.1)	0.01 (0, 0.1)
MCV (fL)	7	90 (87, 94)	90 (86, 94)	90 (88, 94)	91 (87, 96)
Monocytes (x10^9/L)	30	0.67 (0.44, 0.95)	0.68 (0.45, 0.95)	0.59 (0.33, 0.83)	0.61 (0.33, 0.94)
Neutrophils (x10^9/L)	16	5.6 (2.2, 9)	5.6 (2.2, 8.9)	7 (4, 10.3)	5.5 (1.8, 9.1)
Sodium (mmol/L)	9	138 (135, 140)	138 (135, 140)	137 (133, 141)	137 (134, 140)
Urea (mmol/L)	9	6.5 (4.6, 9.7)	6.2 (4.5, 9)	7.4 (5.1, 11)	9.6 (6.6, 15)

(Continued)	% missing	All patients	Discharged or dead	In hospital	ICU admission
B. Validation data	n = 3252	n = 2854	n = 72	n = 326	
Male sex	0	56%	56%	46%	61%
Age (years)	0	71 (58, 80)	69 (55, 79)	72 (58, 82)	79 (73, 85)
BMI (kg/m²)	59	26 (23, 30)	26 (23, 30)	26 (24, 30)	25 (22, 29)
HR (bpm)	40	90 (78, 105)	90 (78, 104)	84 (75, 106)	92 (78, 105)
SBP (mmHg)	43	134 (119, 151)	135 (120, 152)	134 (122, 149)	129 (110, 141)
RR (/min)	43	20 (16, 24)	20 (16, 24)	20 (17, 26)	23 (19, 27)
Saturation (%)	40	95.7 (94, 97.5)	96 (94, 97.7)	95.5 (94, 97)	94.8 (92.3, 96.5)
Temperature (°C)	42	37.3 (36.7, 38.1)	37.3 (36.7, 38.1)	37.3 (36.8, 38.1)	37.2 (36.4, 38)
CRP (mg/L)	9	57 (16, 124)	54 (15, 120)	76 (21, 169)	80 (33, 159)
D dimer (µg/L)	76	1060 (531, 2170)	1013 (490, 2012)	1080 (640, 2570)	1495 (870, 3724)
LDH (U/L)	22	247 (203, 334)	242 (200, 317)	281 (226, 390)	315 (238, 489)
Leucocytes (x10^9/L)	10	9.4 (6.6, 12.9)	9.4 (6.6, 12.9)	9.6 (7, 13.2)	9.5 (6.8, 13.3)
Lymphocytes (x10^9/L)	20	0.98 (0.62, 1.5)	1 (0.64, 1.5)	1.1 (0.72, 1.53)	0.71 (0.48, 1.2)
Albumin (g/L)	20	39 (35, 42)	39 (35, 42)	36 (32, 39)	36 (32, 40)
Bicarbonate (mmol/L)	50	23.5 (21, 26)	23.6 (21, 26)	22.8 (20, 27)	22.8 (20, 25)
Creatinine (µmol/L)	10	84 (66, 116)	83 (65, 111)	80 (58, 118)	103 (74, 158)
Eosinophils (x10^9/L)	27	0.03 (0.01, 0.1)	0.03 (0.01, 0.1)	0.03 (0.01, 0.09)	0.03 (0, 0.04)
MCV (fL)	10	90 (87, 94)	90 (87, 94)	92 (88, 95)	91 (88, 96)
Monocytes (x10^9/L)	30	0.67 (0.43, 0.98)	0.67 (0.44, 0.98)	0.57 (0.38, 1)	0.58 (0.36, 0.97)
Neutrophils (x10^9/L)	21	5.8 (2.4, 9.4)	5.8 (2.4, 9.4)	6.6 (3.3, 9.1)	5.6 (2.2, 9.4)
Sodium (mmol/L)	11	137 (134, 139)	137 (134, 139)	136 (133, 139)	137 (133, 140)
Urea (mmol/L)	11	6.9 (4.9, 10.4)	6.6 (4.7, 9.8)	7.3 (5.3, 11.8)	10.3 (7.4, 17)

Table 5.1: Baseline characteristics of development and validation patient cohorts median ('M') and quartile range ('Q1'=first quartile; 'Q3'=third quartile) are presented for all continuous variables

	% missing	All patients	Discharged or dead	In hospital	ICU admission
A. Development data		n = 2633	n = 2387	n = 32	n = 214
Male sex	0	58%	57%	63%	68 %
Age (years)	0	71 (58, 80)	71 (57, 80)	80 (71, 85)	68 (59, 74)
BMI (kg/m²)	48	26 (23, 30)	26 (23, 30)	25 (22, 29)	27 (24, 31)
HR (bpm)	14	88 (77,100)	88 (77, 100)	87 (77, 98)	91 (79, 104)
SBP (mmHg)	20	131 (116, 149)	131 (116 , 149)	140 (122 , 159)	129 (114 , 145)
RR (/min)	15	19 (15, 23)	19 (15, 23)	18 (16 , 22)	23 (18 , 27)
Saturation (%)	10	95.(94, 97).3	95.(94, 97).4	96 (95, 97).5	95 (93.1, 96).7
Temperature (°C)	11	37.2 (36.7, 37.9)	37.2 (36.7, 37.9)	37.4 (36.7, 37.8)	37.4 (36.7, 38.2)
CRP (mg/L)	4	(11, 120)	47 (10, 110)	82 (41, 175)	88 (20, 178)
D dimer (µg/L)	79	1100 (463, 2700)	1040 (460, 2453)	3400 (2000, 5400)	2000 (590, 4900)
LDH (U/L)	11	239 (197, 317)	234 (194, 306)	270 (216, 324)	318 (234, 444)
Leucocytes (x10^9/L)	4	9.5 (6.7, 13.2)	9.4 (6.8, 13.1)	12 (9.2, 16.1)	10.3 (6.3, 14.7)
Lymphocytes (x10^9/L)	11	1 (0.67, 1.6)	1 (0.7, 1.6)	0.74 (0.55, 1.3)	1.1 (0.7, 1.5)
Albumin (g/L)	8	40 (36, 43)	40 (36, 43)	37 (34, 41)	38 (35, 42)
Bicarbonate (mmol/L)	66	23 (21, 25)	23 (21, 25)	24 (22, 26)	21 (19, 24)
Creatinine (µmol/L)	5	85 (67, 113)	84 (67, 112)	85 (64, 113)	93 (73, 125)
Eosinophils (x10^9/L)	27	0.1 (0.03, 0.1)	0.1 (0.03, 0.11)	0.03 (0.03, 0.1)	0.1 (0.03, 0.1)
MCV (fL)	5	91 (87, 95)	91 (87, 95)	92 (90, 97)	91 (88, 95)
Monocytes (x10^9/L)	41	0.72 (0.47, 1)	0.73 (0.48, 1)	0.71 (0.56, 1.02)	0.64 (0.39, 1)
Neutrophils (x10^9/L)	11	7.2 (4.8, 10.5)	7.1 (4.8, 10.4)	9.8 (6.6, 14.5)	7.7 (5, 11.1)
Sodium (mmol/L)	5	137 (134, 140)	137 (134, 140)	138 (134, 140)	137 (134, 139)
Urea (mmol/L)	5	6.6 (4.7, 9.9)	6.6 (4.6, 9.8)	8.4 (6.3, 11.9)	7.1 (5.1, 11)

(Continued)	% missing	All patients	Discharged or dead	In hospital	ICU admission
B. Validation data		n = 1466	n = 1356	n = 24	n = 86
Male sex	0	58%	58 %	38%	71%
Age (years)	0	71 (57, 80)	71 (56, 80)	80 (69, 85)	70 (62, 76)
BMI (kg/m²)	40	26 (23, 30)	26 (23, 30)	25 (22, 31)	28 (24, 30)
HR (bpm)	16	89 (77, 103)	89 (77, 103)	78 (73, 99)	89 (78, 103)
SBP (mmHg)	21	134 (119, 151)	134 (119, 151)	132 (111, 149)	141 (124, 157)
RR (/min)	17	19 (16, 23)	19 (16, 23)	19 (16, 22)	23 (19, 28)
Saturation (%)	9	95.7 (94, 97.3)	95.9 (94, 97.5)	96.2 (94.5, 97)	93.8 (91.3, 95.7)
Temperature (°C)	13	37.2 (36.7, 37.9)	37.2 (36.7, 37.9)	37.2 (36.7, 37.7)	37.4 (37, 38)
CRP (mg/L)	5	49 (14, 115)	47 (13, 110)	77 (25, 135)	114 (40, 206)
D dimer (µg/L)	95	460 (210, 1275)	440 (210, 1350)	770 (770, 770)	913 (890, 937)
LDH (U/L)	13	238 (196, 314)	234 (195, 307)	254 (217, 288)	360 (234, 531)
Leucocytes (x10^9/L)	6	9.7 (6.8, 13.2)	9.7 (6.8, 13.2)	12 (7.3, 15)	8.9 (6.7, 13.1)
Lymphocytes (x10^9/L)	12	1 (0.65, 1.5)	1 (0.66 , 1.5)	1.1 (0.8, 1.33)	0.8 (0.6, 1.55)
Albumin (g/L)	9	40 (37, 42)	40 (37, 42)	37 (35, 38)	39 (36, 42)
Bicarbonate (mmol/L)	65	23 (21, 25)	23 (21, 25)	23 (21, 26)	22 (19, 25)
Creatinine (µmol/L)	6	86 (68, 119)	85 (68, 117)	80 (58, 122)	96 (77, 138)
Eosinophils (x10^9/L)	20	0.07 (0.03, 0.1)	0.07 (0.03, 0.11)	0.05 (0.03, 0.1)	0.05 (0.03, 0.1)
MCV (fL)	6	91 (88, 95)	91 (88, 95)	94 (90, 97)	91 (87, 95)
Monocytes (x10^9/L)	33	0.7 (0.47, 1)	0.7 (0.48, 1)	0.68 (0.48, 1.03)	0.61 (0.4, 1)
Neutrophils (x10^9/L)	13	7.4 (4.8 , 11)	7.3 (4.7, 11)	9 (5.5, 13.1)	7.4 (5 , 10.8)
Sodium (mmol/L)	6	137 (134, 139)	137 (134, 139)	136 (133, 138)	135 (132, 138)
Urea (mmol/L)	7	6.7 (4.9 , 10)	6.6 (4.8, 9.9)	8 (4.7, 12.8)	7.8 (5.9, 12.2)

Table 5.2: Baseline characteristics of development and validation patient cohorts of two hospitals with a full record of ICU admissions median ('M') and quartile range ('Q1'=first quartile; 'Q3'=third quartile) are presented for all continuous variables

		Univariable			Full model			Selected model			COPE
Predictor	Contrast	OR	95% CI	Wald	OR	95% CI	Wald	OR	95% CI	Wald	OR
Month	≥ April vs ≤April	2.57	2.16 to 3.05	114	1.99	1.61 to 2.47	39	2.06	1.68 to 2.52	49	1.96
Sex	male vs female	1.37	1.15 to 1.63	13	1.12	0.9 to 1.39	1				
Age (years)	80 vs 58	3.07	2.63 to 3.58	201	3.16	2.56 to 3.91	113	2.95	2.42 to 3.6	112	2.74
BMI (kg/m)	35 vs 25	1.07	0.89 to 1.28	1	1.09	0.9 to 1.34	1				
HR (bpm)	103 vs 78	1.16	1.05 to 1.29	8	1.19	0.97 to 1.45	3				
SBP (mmHg)	150 vs 118	0.76	0.65 to 0.89	12	0.86	0.73 to 1.01	3				
RR (/min)	23 vs 16	1.98	1.77 to 2.21	150	1.63	1.34 to 1.99	24	1.91	1.63 to 2.23	64	1.82
Saturation (%)	97.5 vs 94	0.61	0.52 to 0.72	37	0.77	0.65 to 0.9	10				
Temperature (C)	38 vs 37	1.16	1.02 to 1.32	5	1.05	0.87 to 1.27	0				
CRP (mg/L)	118 vs 10	2.76	2.35 to 3.25	149	1.54	1.22 to 1.93	14	1.57	1.27 to 1.93	18	1.52
LDH (U/L)	322 vs 200	2.17	1.99 to 2.36	309	1.83	1.62 to 2.06	99	1.85	1.66 to 2.05	127	1.77
Leucocytes (x10^9/L)	12.7 vs 6.7	1.01	0.92 to 1.11	0	0.88	0.75 to 1.02	3				
Lymphocytes (x10^9/L)	1.6 vs 0.66	0.67	0.6 to 0.75	48	1.03	0.9 to 1.19	0				
Albumin (g/L)	42 vs 36	0.58	0.53 to 0.62	191	0.8	0.71 to 0.9	14	0.77	0.69 to 0.86	21	0.78
Bicarbonate (mmol/L)	25.9 vs 21.4	0.71	0.65 to 0.78	54	0.81	0.71 to 0.92	10				
Creatinine (Ámol/L)	111 vs 66	1.58	1.46 to 1.71	133	0.85	0.72 to 1	4				
Eosinophils (x10^9/L)	0.1 vs 0.004	0.77	0.64 to 0.93	7	1.08	0.9 to 1.3	1				
MCV (fL)	94 vs 87	1.18	1.08 to 1.29	14	1.1	0.99 to 1.23	3				
Monocytes (x10^9/L)	0.95 vs 0.44	0.84	0.74 to 0.96	7	1.08	0.87 to 1.35	1				
Neutrophils (x10^9/L)	9 vs 2.2	0.97	0.88 to 1.07	0	0.94	0.82 to 1.08	1				
Sodium (mmol/L)	140 vs 135	0.98	0.96 to 1.01	1	1.03	0.98 to 1.09	1				
Urea (mmol/L)	9.7 vs 4.6	2.48	2.24 to 2.76	291	1.79	1.43 to 2.24	26	1.61	1.42 to 1.83	53	1.56

Table 5.3: Univariable and multivariable associations between predictors and death within 28 days OR with 95for a model with all available predictors (columns 'full model') and for a model with only the six strongest predictors (columns 'selected model')

		Univariable			Multivariable		
Predictor	Contrast	OR	95% CI	Wald	OR	95% CI	Wald
Month	≥April vs ≤April	2.06	1.51 to 2.81	21	1.63	1.16 to 2.28	8
Age (years)	80 vs 58	1.96	1.47 to 2.62	21	1.76	1.32 to 2.35	15
RR (/min)	23 vs 16	1.76	1.48 to 2.09	40	1.71	1.4 to 2.09	27
CRP (mg/L)	118 vs 10	1.88	1.44 to 2.44	22	1.3	0.95 to 1.77	3
LDH (U/L)	322 vs 200	1.73	1.52 to 1.98	66	1.44	1.25 to 1.67	24
Albumin (g/L)	42 vs 36	0.75	0.64 to 0.88	13	0.95	0.78 to 1.17	0
Urea (mmol/L)	9.7 vs 4.6	1.29	1.08 to 1.54	8	1.36	1.1 to 1.66	8
Adjusted for:							
Max[Age-70, 0] (years) *	80 vs 58	0.2	0.13 to 0.3	57	0.17	0.11 to 0.27	65

Table 5.4: Multivariable associations between predictors and ICU admission within 28 days OR with 95variables (columns 'univariable') and for a model with the six strongest predictors of death, corrected for a decreasing probability of ICU admission after the age of 70 (columns 'multivariable')

Discussion

We developed COPE for prediction of in-hospital death and need for intensive care when patients with suspected COVID-19 present at the ED. Developed using patient data from the first wave of the pandemic, based on six quickly and objectively obtainable predictors when entering the ED—age, RR, LDH, CRP, albumin and urea—COPE discriminated well and was well calibrated in patients admitted to hospitals in the second wave of the pandemic, both for predicting in-hospital death and for ICU admission.

The clinical presentation of COVID-19 is broad and varies from asymptomatic to critical disease. Some patients who initially have mild symptoms progress to severe disease within 1 week.[17]In the ED physicians need to identify high-risk patients—that is, those at high risk of deterioration and/or death—requiring treatment in the ICU, intermediate-risk patients requiring admission to the clinical ward, and low-risk patients who can potentially be sent home. Since COPE is based on data that are routinely measured, or at least readily available in the ED, it can act as a tool to support such decisions. Hospitalised patients who are at high risk for mortality or need for ICU admission should be more intensively watched, and when a high load of high-risk patients occurs in the ED, this should be taken into account in the ICU capacity planning. COPE does not explicitly define treatment decisions based on risk thresholds, such as: send the patient home when the mortality risk is below a risk threshold of x%, or: admit the patient to the ICU over a mortality risk threshold of y%. These currently unavailable risk thresholds, and hence the resulting treatment decisions, depend on a trade-off between benefits and harms (including costs) of hospital or ICU admission.[18] Further research is necessary to better understand the benefits and harms of hospital admission and of ICU admission, for individual patients with COVID-19.[19] Furthermore, treatment decisions may depend on the availability of resources. The decision to admit a patient to the hospital, or even to the ICU, may depend on the availability of hospital beds and ICU beds. Likewise, the decision to send a low-risk patient home may depend on the availability of relatives who are willing to care for the patient at home. Although it is currently not possible to define explicit risk-based treatment decisions for patients with COVID-19, the risk predictions provided by COPE can be factored in by doctors, patients and relatives, when making decisions about hospital or ICU admission.

CU admission. We requested 19 large Dutch hospitals to supply anonymised retrospective data on the cohorts of patients with COVID-19 who were admitted to their hospital. This request for data was sent out very early in the pandemic and was greeted with enthusiasm. Probably due to the enormous

pressure on healthcare at that time, four hospitals supplied useable data for the analysis. The contributing hospitals were well spread over the Netherlands, with one in the west, two in the south and one in the east of the country and are a mix of academic and large teaching hospitals. we believe they are representative for healthcare in the Netherlands. Although the consistently good performance of COPE across the hospitals may support its generalisability to other countries, geographical validation would be additionally reassuring, since the epidemic, and clinical practice—for example, access to ICU or other enhanced care—for this novel disease, may have substantial intercountry variability.

COPE was developed based on 5831 patients of whom 629 died within 28 days. This effective sample size of 629 events was ample to start the development process with a full model of 45 regression coefficients (14 events per variable), that is, one binary predictor (sex) and 22 continuous predictors with 2 regression coefficients— due to using non-linear terms—each.[20] To prevent too extreme predictions of COPE in new data, we applied a shrinkage factor to its regression coefficients, based on a bootstrap procedure with backward selection starting from the full model.[12]

Our explicit aim was to develop a score based on quickly and objectively obtainable predictors at presentation at the ED. Consequently, pre-existing comorbidities, the level of consciousness measured by the Glasgow Coma Scale, and chest radiography results—although predictive for outcomes of patients with COVID-19 in other studies—were not considered here.[5 7 8] Some predictors were promising in univariable analysis, such as lymphocytes and creatinine, but had negligible effects in multivariable analysis, because of strong correlations with other, more important predictors. Other predictors, such as oxygen saturation and bicarbonate, were significantly associated with death in multivariable analysis, but were not selected into the final model, since our explicit aim was to develop a simple model and the incremental value of these predictors was minimal. To achieve this aim, we only selected the strongest predictors—age, RR, LDH, CRP, albumin and urea—resulting in a parsimonious but well-performing model.

We aimed to predict outcomes for all patients who present to the ED with suspected COVID-19, regardless of actual hospital admission. Our data were limited to patients who presented at the ED and were admitted to hospital, because their outcomes were captured in the retrospective hospital database, while outcomes of patients who were sent home were not captured in the retrospective hospital database. Nevertheless, over 90% of the patients who presented to the ED with suspected COVID-19 were admitted to hospital and it is reasonable to assume that our predictions can be extrapolated to the less

than 10% of patients who were sent home. Of note, the discriminative ability of our model is probably better in all patients presenting to the ED, due to a more heterogeneous casemix: patients who were sent home are likely to have more favourable predictor levels and more favourable outcomes than patients who are admitted.[14]

Besides mortality, we aimed to predict the need for ICU admission. A limitation of our study is that the need for ICU admission differs from the observed decisions on ICU admission, and is inherently difficult to model, because recorded ICU admissions express historical decisions at national, regional, hospital or even intensivist level. As a robust solution, we exploited the strong correlation between need for intensive care and death, by calibrating our model for predicting death to the observed ICU admissions, adjusting for a linear decrease with age after the age of 70. Hence, we assumed a linear relationship between (the logarithm of the odds of) death and need for ICU admission, and that all patients below the age of 70 needing intensive care were actually admitted to the ICU, that is, the need for ICU admission is well estimated by the observed decisions on ICU admission for patients below the age of 70. The latter is reasonable given the sufficiency of ICU beds for Dutch patients throughout the pandemic. The discriminative ability of this recalibration approach was very similar to that of a model that refitted all associations between COPE predictors and ICU admission. With temporal validation in two separate hospitals, we showed that COPE discriminated very well between patients at low and high risk of ICU admission and that the predicted probability of ICU admission was well calibrated for the 20% highest-risk patients (highestrisk quintiles in ***figure 3***). Nevertheless, recalibration of COPE for predicting need for ICU admission to local circumstances may be necessary.

The absence of external validation in our study— measuring the predictive performance of COPE in hospitals that were not present in the development data—may be considered a limitation of this study.[21] However, the combination of temporal validation—in second wave patients—and geographical validation—in separate hospitals—is a strength of this study.[22] Although COPE already performed very well when validated across time and space, future research should focus on analyses of potential time trends not captured by the predictors—for example, changes in mortality due to: improvements in treating patients with COVID-19; mutations of COVID19; changes in patient casemix or critical care capacity fluctuations [23]—potential changes in predictor effects in time (interactions between predictors and time), and the impact of potential differences in patient case mix and differences in clinical care in countries other than the Netherlands (international validation). These casemix and clinical care differences should primarily affect calibration, requir-

ing an update of the model intercept, but not discrimination. The emergence of new COVID-19 variants with potentially different mortality risk may especially require frequent analyses of the need for model updating.[24]

In conclusion, COPE, a simple tool based on six quickly and objectively obtainable predictors in the ED, is well able to predict mortality and need for ICU admission for patients who present to the ED with suspected COVID19. COPE may support patients and doctors in decision making.

Predictor	Minimum	Maximum
Age (years)	0	100
RR (/min)	10	60
CRP (mg/L)	1	500
LDH (U/L)	50	4000
Albumin (g/L)	10	60
Urea (mmol/L)	1	80
$\text{lp} = -13.6 + 0.04575 \times \text{age} + 1.654 \times \log(\text{RR}) + 0.1688 \times \log(\text{CRP}) + 1.197 \times \log(\text{LDH}) + -1.585 \times \log(\text{albumin}) + 0.5953 \times \log(\text{urea})$		
Probability of death within 28 days $= \frac{1}{(1+exp(-\text{lp}))}$		
Probability of ICU admission within 28 days $= \frac{1}{(1+exp(-(-0.08949+0.5970\times\text{lp})))}$		

Table 5.5: COPE definition. COPE: COVID-19 outcome prediction in the emergency department; CRP: C reactive protein; ICU: intensive care unit; LDH: lactic dehydrogenase: RR: respiratory rate.

CHAPTER 6

Treatment heterogeneity in the study of the comparative effectiveness of teriparatide vs bisphosphonates in routine practice conditions: a multi-database cohort study

Abstract

Objectives: To study the comparative effectiveness of teriparatide (TP) vs oral bisphosphonates (BP) to reduce hip, major osteoporotic and vertebral fracture risk. In addition, we stratified by predicted hip fracture risk to assess treatment effect heterogeneity.

Materials and Methods: We conducted a network cohort study using data from four US-based databases, namely IBM MarketScan® Commercial Claims and Encounters (CCAE), IBM MarketScan® Medicare Supplemental Beneficiaries (MDCR), Optum®De-Identified Clinformatics Data Mart Database – Date of Death (OPTUM-DOD) and Optum® de-identified Electronic Health Record Dataset (OPTUM-EHR), all mapped to the OMOP common data model. We included all women aged >50, who initiated TP or BP and had no history of anti-osteoporotic treatment in the prior year. Propensity scores were used for 1:4 matching to minimise confounding by indication. Models to predict hip fracture risk were developed and validated separately in each of the four databases. Finally, 147 negative control outcomes (NCO) were included to calibrate for residual confounding. Cox regression was used to estimate calibrated hazard ratios (HR) and Kaplan-Meier estimated differences 3 years after treatment initiation to estimate absolute effects. We provide meta-analytic estimates for the overall analysis and for patients below and above 3% hip fracture risk.

Results: A total of 35,869 and 133,437 users of TP and BP contributing 75,649 and 280,091 person years, respectively, were included from all four databases. NCO analyses showed evidence of residual confounding, hence we report empirically calibrated estimates: Overall meta-analytic HR were 0.87 (0.74 to 1.02; 95% CI), 1.10 (1.02 to 1.18) and 1.10 (1.00 to 1.21; 95% CI) for hip fracture, major osteoporotic fracture and vertebral fracture respectively. Meta-analytic HR for patients below 3% hip fracture risk were 1.00 (0.85 to 1.18; 95% CI), 1.24 (1.10 to 1.39; 95% CI), and 1.24 (1.07 to 1.43;95% CI), respectively. In patients at hip fracture risk above 3% the respective estimates were 0.83 (0.64 to 1.07; 95% CI), 0.86 (0.73 to 1.01; 95% CI), and 1.04 (0.86 to 1.26; 95% CI).

Conclusions: Overall, we found negligible differences in comparative fracture prevention effectiveness of TP vs BP. However, our study suggests relevant treatment effect heterogeneity, with a tendency towards favouring TP in patients with high anticipated hip fracture risk.

Introduction

Osteoporosis is a chronic condition characterised by decreased bone density and increased risk for fragility fractures that affects almost 30% of women aged 50 years[165]. In 2017, over 2.7 million incident fragility fractures occurred in the 5 largest EU countries and Sweden, with hip fractures accounting for 19.6% of fractures but up to 57% of total costs[166]. By 2030, annual fractures are expected to rise by almost 23%, with related costs increasing by 27%[166]. Fracture prevention is thus the key focus of anti-osteoporotic therapy. Several pharmacological agents are available, with the choice of anti-osteoporotic agent largely depending on fracture history and anticipated fracture risk. Oral bisphosphonates (BP) are first-line treatments for postmenopausal women with increased fracture risk, considering their favourable cost-effectiveness and safety profile. Anabolic agents are therefore spared for people who do not respond to BP or who are at very high fracture risk. Teriparatide (TP), a parathyroid-hormone analogue administered as daily injections, was the first anabolic agent approved by the FDA for the treatment of severe postmenopausal osteoporosis.

Clinical trials have shown that teriparatide is efficacious to substantially reduce vertebral fractures compared to placebo[167] and risedronate[168]. However, its efficacy on low-incident osteoporotic fractures, especially hip fractures, is less well established. Previous randomised trials assessing hip fracture comprised only few events, thus not providing sufficient power for comparative effectiveness analyses. A recent meta-analysis indicates a significant 80% reduction in risk for hip fracture with teriparatide compared to placebo and a non-significant 46% risk reduction when compared to active controls[169].

Clinical effectiveness in routine practice may differ from findings in clinical trials due to many reasons including poor compliance[170], and treatment benefit may largely depend on patient's underlying fracture risks. We therefore leveraged multinational large real-world data from electronic medical records and health claims to study the comparative effectiveness of teriparatide vs bisphosphonates in actual practice conditions. Additionally, we used novel methods[101] to test for treatment heterogeneity according to baseline fracture risk.

6.1 Methods

Study population

We performed a new user cohort study to estimate the effectiveness of TP compared to BP in patients with osteoporosis[170]. New users were defined based on no history of anti-osteoporosis drugs use in the 365 days prior to treatment initiation with TP or an BP. We included female participants above

the age of 50 with established osteoporosis defined by a history of hip, wrist, spine or shoulder/humerus fracture prior to treatment initiation. We excluded patients with less than 1 year of data available before treatment start. More information on cohort definitions are available in section 1 of the supplement and in **supplementary Tables S1-S10**.

Study design

Our primary effectiveness outcome was hip fracture, as this is typically the fracture outcome most reliably recorded in the proposed datasets. Vertebral fracture and a composite major osteoporotic fracture, defined as hip, vertebral or wrist/forearm/proximal humerus fracture, were our secondary effectiveness outcomes Patients were followed for a maximum of three years (1095 days) after treatment initiation. Continuous exposure to the treatments under study was achieved by imposing a maximum of 30 days between prescriptions. In cases where this was not the case, patients were censored upon treatment discontinuation, that is, 30 days after their last prescription. We also censored patient follow-up in cases of drop out from the database, death or at the time when anti-osteoporotic treatment was switched.

Data sources

We ran our analyses on four US observational databases mapped to OMOP-CDM version. More specifically:

- IBM MarketScan Commercial Database (CCAE) includes health insurance claims across the continuum of care (e.g. inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare) as well as enrollment data from large employers and health plans across the United States who provide private healthcare coverage for more than 155 million employees, their spouses, and dependents. This administrative claims database includes a variety of fee- for-service, preferred provider organizations, and capitated health plans.
- IBM® MarketScan® Medicare Supplemental Database (MDCR) represents the health services of approximately 10 million retirees in the United States with Medicare supplemental coverage through employer-sponsored plans. This database contains primarily fee-for-service plans and includes health insurance claims across the continuum of care (e.g. inpatient, outpatient and outpatient pharmacy).
- Optum De-Identified Clinformatics® Data Mart Database - Date of Death (Optum-DOD) is derived from a database of administrative health claims for members of large commercial and Medicare Advantage health plans. The database includes approximately 17-19 million annual

covered lives, for a total of over 65 million unique lives over a 12 year period (1/2007 through 12/2019). The population is geographically diverse, spanning all 50 states.

- Optum de-identified Electronic Health Record Dataset (Optum-EHR) is derived from dozens of healthcare provider organizations in the United States, that include more than 700 Hospitals and 7000 Clinics; treating more than 102 million patients receiving care in the United States.

Statistical analyses

Comparative effectiveness

We derived overall treatment effect estimates of TP compared to BP regarding the three outcomes of interest using Cox proportional hazards models with treatment as the sole covariate. We adjusted for observed confounding by fitting the Cox regression models on the propensity score matched (1:4) subset of the study population. Hazard ratios (HR) derived in different databases were summarized using random effects meta-analysis (**REF**).

We developed separate large-scale propensity score models within each database using LASSO logistic regression with a predefined set of measured covariates[171]. We qualitatively assessed the ability of the estimated propensity scores to adjust for observed confounding using the following diagnostics. First, we evaluated the overlap between TP and BP treated patients by plotting the distributions of the preference scores, that is, a transformation of the propensity score that adjusts for prevalence differences between treatment arms. A common rule of thumb requires the majority of the patients to lie between 0.3 and 0.7 for both treatment arms to assume that patients are comparable[172]. Second, we evaluated covariate balance before and after propensity score adjustment by plotting the absolute standardized differences (SMD) of all measured covariates before and after matching on the propensity scores. A commonly used rule of thumb requires SMD to be less than 0.1 in order to assume adequate covariate balance after propensity score adjustment[25,173,174].

Residual confounding

Residual study bias from unmeasured confounding can still be present in observational studies, which may not be visible with standard diagnostics. Negative control outcome (NCO) analyses are a popular approach for assessing the presence of residual confounding[175]. Negative controls are treatment-outcome pairs where a null treatment effect has been established. Using the effect estimates in a set of 126 NCO experiments, i.e. outcomes that are not caused

by TP or BP, we derived an empirical approximation to the null distribution. We used this approximation to calibrate all estimated hazard ratios and their 95% confidence intervals for the three outcomes of interest[27,29].

Treatment effect heterogeneity

We assessed heterogeneity of treatment effects across patient groups with varying baseline risk using the recently proposed framework for assessment of treatment effect heterogeneity in observational data[16,17,176]. We implemented the method as follows for our study: we first derived individualized risk predictions for the three fracture outcomes. We built the prediction models using LASSO logistic regression on the propensity score matched (1:10) subpopulation of the pooled treatment arms. We then considered the same large set of candidate covariates as for the development of the propensity score models.

For each outcome we used the derived prediction models for hip fracture risk to stratify the patients into two risk groups: patients below 3% and patients above 3% three-year hip fracture risk. Within each of these risk groups we developed a new propensity score model. Our analyses were performed on the propensity score matched (1:4) subset of the risk subgroup subset. We derived relative effect estimates using Cox proportional hazards models only with treatment as a predictor. Absolute effect estimates were calculated based on the difference of the Kaplan-Meier estimates, on day 1095 after treatment initiation. In the supplement (Supplementary Figures 3-5) we present a second approach where the population is stratified using existing guidelines based on age-dependent 10-year major osteoporotic fracture risk[2].

Results

After propensity score matching, a total of 35,869 TP users contributed 75,849 person-years, with 846 fractures observed. This compared to 133,437 matched BP users, 280,901 person-years, and 3,409 fractures. Before propensity score matching, TP users were more likely to have a diagnosis of osteoarthritis, rheumatoid arthritis, depressive disorder, and gastroesophageal reflux disease compared to BP users (Supplementary Tables 10-13). After propensity score matching these imbalances disappeared (Table 1). In all databases adequate equipoise of the preference score distributions was achieved (Figure 1), with all baseline covariates (>35,000 in each database) well balanced after matching (Figure 2). This indicates that we were able account for observed confounding present in all databases. However, NCO analyses suggested the presence of substantial residual confounding, with matched TP users showing a higher risk of multiple causally unrelated (negative control) outcomes, as shown in Figure 3. This suggests that propensity score matching did not fully account

for confounding by indication, as TP users appear less healthy than matched BP users in terms of many unrelated comorbidities.

After propensity score matching, a total of 35,869 TP users contributed 75,849 person-years, with 846 fractures observed. This compared to 133,437 matched BP users, 280,901 person-years, and 3,409 fractures. Before propensity score matching, TP users were more likely to have a diagnosis of osteoarthritis, rheumatoid arthritis, depressive disorder, and gastroesophageal reflux disease compared to BP users (Supplementary Tables 10-13). After propensity score matching these imbalances disappeared (Table 1). In all databases adequate equipoise of the preference score distributions was achieved (Figure 1), with all baseline covariates (>35,000 in each database) well balanced after matching (Figure 2). This indicates that we were able account for observed confounding present in all databases. However, NCO analyses suggested the presence of substantial residual confounding, with matched TP users showing a higher risk of multiple causally unrelated (negative control) outcomes, as shown in Figure 3. This suggests that propensity score matching did not fully account for confounding by indication, as TP users appear less healthy than matched BP users in terms of many unrelated comorbidities.

We estimated meta-analytic HR of 1.06 (0.92 to 1.22; 95% CI), 1.30 (1.18 to 1.44; 95%CI), and 1.30 (1.14 to 1.49; 95% CI) for hip fracture, major osteoporotic fracture, and vertebral fracture, respectively, in the lower hip fracture risk group across the four databases. Using the results of the NCO analyses, we recalibrated these estimates to 1.00 (0.85 to 1.76; 95% CI), 1.24 (1.10 to 1.39; 95% CI), and 1.24 (1.07 to 1.43; 95% CI) for hip fracture, major osteoporotic fracture, and vertebral fracture, respectively. As for patients at higher hip fracture risk, corresponding meta-analytic HRs were 0.91 (0.75 to 1.11; 95% CI), 0.92 (0.76 to 1.10; 95% CI) , and 1.11 (0.93 to 1.34; 95% CI), respectively, which were recalibrated to 0.83 (0.64 to 1.07; 95% CI), 0.86 (0.73 to 1.01; 95% CI), and 1.04 (0.86 to 1.26; 95% CI), respectively. This difference in effect indicated an interaction between hip fracture risk and the comparative effectiveness of TP vs BP treatment, with a potential better protection in higher fracture risk. Risk stratified negative control analyses, again, showed evidence of residual confounding (Supplementary Figure 2).

The stronger relative treatment effects with increasing three-year hip fracture risk found in MDCR and Optum-DOD translated to increasing benefits on the absolute scale. Absolute treatment effect estimates (risk differences) increased from -0.03% (-0.68 to 0.62%; 95% CI) to 1.55% (0.19% to 2.91%; 95% CI) and from 0.19% (-0.23% to 0.62%; 95% CI) to 1.71% (0.37% to 3.06%; 95% CI) in MDCR and Optum-DOD respectively. No similar differences on the absolute scale were observed in Optum-EHR. The majority of the patients in CCAE

were below 3% three-year hip fracture risk, therefore, we only present results for the lower risk group. We found negligible difference between TP and BP on the absolute scale in CCAE for patients below 3% three-year hip fracture risk, which further supports the similarity of the compared treatments for lower risk patients. For patients below 3% three-year hip fracture risk, we actually found absolute risk increase with TP in CCAE and Optum-EHR, with absolute effect estimates of -0.98% (-1.61% to -0.36%; 95% CI) and -1.09% (-1.71% to -0.46%; 95% CI), respectively. For patients above 3% three-year hip fracture risk we found absolute benefits with TP treatment in Optum-EHR (3.20% with 95% CI from 1.07% to 5.33%). Finally, we found vertebral fracture risk increase with TP treatment in patients below 3% three-year hip fracture risk of -0.67% (-1.15% to -0.19%; 95% CI) and 1.11% (-1.69% to -0.53%; 95% CI) in CCAE and Optum-HER, respectively. We found negligible differences for vertebral fracture risk between treatments in patients above 3% three-year hip fracture risk.

6.2 Discussion

6.2.1 Key findings

Overall, we found negligible differences in comparative fracture prevention effectiveness of TP vs BP. However, we demonstrated relevant treatment heterogeneity, with a tendency towards improved anti-fracture effectiveness with teriparatide among patients with higher hip fracture risk. Negative control outcome analyses suggested the presence of unresolved confounding. We used empirical calibration to correct for unresolved confounding, which moved our treatment effect estimates closer to those observed in previous randomised controlled trials and systematic reviews.

6.2.2 Effectiveness and treatment heterogeneity

This is the first large-scale observational study assessing treatment heterogeneity in the effectiveness of TP and BP for hip fracture prevention in postmenopausal women. Previous randomised trials assessing the efficacy of TP for preventing hip fractures included only few events, and thus didn't allow for stratification based on fracture risk. According to the Postmenopausal Osteoporosis Guideline issued by the American Association of Clinical Endocrinologists/American College of Endocrinology 2020, TP should be considered as initial therapy for patients at very high fracture risk, or individuals who are unable to use oral anti-fracture therapy[177]. Thus, TP is typically reserved for patients suffering from severe osteoporosis due to higher treatment costs compared to BP. However, with TP becoming available as a generic drug, costs for treatment are dropping, potentially affecting cost-effectiveness estimates and

future related guidelines. Results from our study suggest that TP treatment could indeed be more effective for patients with higher hip fracture risk. However, this will need to be evaluated further in future studies as the observed risk reduction was not significant in our study.

6.2.3 Findings in the light of residual confounding

Unlike randomized controlled trials, treatment allocation is not random in observational data. While preference score distributions showed a substantial overlap, patients receiving TP or BP likely differed in their baseline fracture risk. Our negative control outcome analyses showed that these differences were not completely balanced after PS matching, and unmeasured confounding remained.

Similar to our study, previous studies also found residual confounding when comparing oral BPs users with patients initiating parenteral anti-osteoporosis treatments[178]. Users of parenteral treatments had more contact with the healthcare system and higher comorbidity burden compared to BP users, and these differences could not be removed completely by PS weighting. This highlights the complexity of comparing parenteral and oral anti-osteoporosis therapies.

We used empirical calibration to correct for unresolved confounding, which subsequently moved our treatment effect estimates closer to those observed in previous trials: the Vero trial reported a reduced risk for clinical fractures associated with TP used compared to risedronate in post-menopausal women with severe osteoporosis (HR 0.48 [0.32–0.74]). While our study shows a lot of uncertainty and differences between databases, estimates from calibrated meta-analyses for major osteoporotic fractures are closer to trial estimates in patients with high hip fracture risk (HR 0.86 [0.73 to 1.01]), compared to people with low fracture risk. Moreover, our results from calibrated meta-analyses for hip fracture risk (HR 0.87 [0.74 to 1.02]) are comparable to findings from a recent meta-analysis, which reported a non-significant 46% risk reduction for hip fractures (OR 0.44 (0.22-0.87) compared to active controls (mostly oral BP in 10/20 studies)[169].

6.2.4 Study strengths and limitations

This study has multiple strengths and limitations. Using multiple large claims databases, enough outcome events were identified during follow-up to ensure statistical power, with Minimum Detectable Rate Ratios (MDRR) between 1.2 and 1.4 for hip fracture across databases. While hip fracture is a comparatively rare outcome, it is unambiguously defined and reliably recorded in routinely collected data. Major osteoporosis fracture was studied allowing for a direct

comparison to results from RCT.

While our study used sophisticated prediction models to stratify for fracture risk, clinical measures such as bone mineral density typically used for fracture risk assessment was not available in our data and could therefore not be considered. Results from previous non-controlled, observational studies found reduced rates of hip fractures in patients with high vs. low teriparatide adherence[179] and for longer treatment (>12 months) compared to the first 6 months after teriparatide initiation[180]. Our study did not consider adherence, but only censored follow up in case of treatment cessation or a large treatment gap, and median follow-up for our study was only 6 months.

6.2.5 Conclusions

Our study found relevant treatment heterogeneity, with a tendency towards favouring teriparatide in patients with high anticipated hip fracture risk. While we approximated trial findings, our study seems to systematically underestimate the effectiveness of teriparatide likely due to unresolved confounding.

	CCAE			MDCR			OPTUM-DOD			OPTUM-EHR		
	TP (%)	BP (%)	Std. diff	TP (%)	BP (%)	Std. diff	TP (%)	BP (%)	Std. diff	TP (%)	BP (%)	Std. diff
Medical history: General												
Attention deficit hyperactivity disorder	1.2	1.2	0	0.3	0.2	0.02	0.8	0.9	-0.01	0.6	0.6	0
Chronic obstructive lung disease	10.8	10.9	-0.01	23.9	25.3	-0.03	22	23.2	-0.03	16.6	17.1	-0.01
Crohn's disease	1.8	1.8	0	1.3	1.2	0.01	1.3	1.4	-0.01	1.2	1.1	0.01
Depressive disorder	16	17.1	-0.03	14.3	15.5	-0.03	20.4	22.1	-0.04	21.2	21.7	-0.01
Diabetes mellitus	0.2	0.3	-0.01	0.1	0.1	0	0.3	0.4	-0.02	1.4	1.4	0.01
Gastroesophageal reflux disease	26.4	27.5	-0.02	31.8	34.5	-0.06	31	32.9	-0.04	30	29.9	0
Gastrointestinal hemorrhage	3.3	3.4	-0.01	6.7	7.1	-0.01	5.6	5.7	-0.01	3.5	3.6	0
Human immunodeficiency virus infection	0.2	0.2	-0.01	0	0	-0.02	0.2	0.2	0.01	0.1	0.2	-0.02
Hyperlipidemia	39.7	40.6	-0.02	46.4	48.8	-0.05	54.4	55.7	-0.03	43.4	42	0.03
Obesity	6.8	7.5	-0.03	5.6	6.1	-0.02	11.5	12.3	-0.02	10.1	10.2	0
Osteoarthritis	14.3	14.5	-0.01	25.3	26.9	-0.04	29	30.1	-0.02	23.2	23.1	0
Pneumonia	1.6	1.9	-0.02	2.4	2.4	0	4.5	4.9	-0.02	5	5.3	-0.01
Psoriasis	2.6	2.6	0	2.6	3	-0.03	2.6	2.8	-0.02	1.6	1.7	-0.01
Rheumatoid arthritis	9.8	10.2	-0.01	11.1	11.4	-0.01	13	13.5	-0.01	10.2	10.7	-0.01
Ulcerative colitis	1.7	1.6	0.01	1.8	1.7	0	1.7	1.6	0.01	1.1	1.1	0
Urinary tract infectious disease	31.9	32.7	-0.02	36.2	38.2	-0.04	39	40.9	-0.04	24.7	24.8	0
Medical history: Cardiovascular disease												
Atrial fibrillation	3	3.1	0	14.6	15.1	-0.01	9.5	10.1	-0.02	9.6	9.8	-0.01
Cerebrovascular disease	1.8	1.8	0	7.8	8.1	-0.01	5	5.3	-0.01	2.6	2.7	-0.01
Coronary arteriosclerosis	5.4	5.4	0	17.8	18.3	-0.01	10.2	10.7	-0.02	8	7.9	0
Heart disease	1.8	1.6	0.01	4	4.3	-0.02	4.3	4.4	-0.01	4.1	4.1	0
Heart failure	1.1	1.1	0	2.7	3	-0.02	4.9	5.3	-0.02	3.5	3.7	-0.01
Ischemic heart disease	0.2	0.2	0	0.6	0.7	-0.01	0.4	0.4	-0.01	0.3	0.2	0.02
Pulmonary embolism	0.4	0.4	0	0.5	0.6	-0.01	1	1.1	-0.01	1.3	1.4	-0.01
Venous thrombosis	1	1.2	-0.01	1.5	1.9	-0.03	1.2	1.4	-0.01	0.8	0.9	-0.01
Medical history: Neoplasms												
Malignant lymphoma	0.5	0.5	0	0.8	1	-0.02	0.3	0.5	-0.03	0.5	0.6	-0.01
Malignant neoplastic disease	0.4	0.4	0	0.6	0.6	-0.01	0.3	0.5	-0.02	0.7	0.8	-0.01

Treatment	Database	Patients	Follow-up time (years)	Events (n)
Teriparatide	CCAE	8,258	14,661	82
	MDCR	6,378	12,575	228
	OPTUM-DOD	8,958	18,809	246
	OPTUM-EHR	12,275	29,804	290
Oral bisphosphonates	CCAE	31,194	55,126	275
	MDCR	23,281	45,855	1,023
	OPTUM-DOD	34,298	71,578	1,141
	OPTUM-EHR	44,664	107,532	970

Figure 6.1: test

```
Warning in
tiff::readTIFF(here::here("figures/ch7-overallNcPlot_itt_att_1095_custom.tif
TIFFReadDirectory: Sum of Photometric type-related color channels and
ExtraSamples doesn't match SamplesPerPixel. Defining non-color channels as
ExtraSamples.
```

```
grid::grid.raster(tiff::readTIFF(here::here("figures/ch7-plotAbsoluteRisk
```

Outcome	Database	Uncalibrated HR	Calibrated HR
Hip fracture	CCAE	1.12 (0.87, 1.43)	1.02 (0.76, 1.38)
	MDCR	0.81 (0.70, 0.93)	0.76 (0.61, 0.94)
	OPTUM-DOD	0.82 (0.71, 0.94)	1.12 (1.01, 1.23)
	OPTUM-EHR	1.08 (0.95, 1.23)	1.04 (0.83, 1.31)
	Meta-analysis	0.94 (0.79, 1.12)	0.87 (0.74, 1.02)
Major osteoporotic fracture	CCAE	1.18 (0.97, 1.44)	1.12 (0.88, 1.41)
	MDCR	1.02 (0.89, 1.18)	1.00 (0.87, 1.15)
	OPTUM-DOD	1.19 (1.07, 1.31)	1.12 (1.01, 1.23)
	OPTUM-EHR	1.10 (0.98, 1.23)	1.09 (0.91, 1.30)
	Meta-analysis	1.15 (1.04, 1.27)	1.10 (1.02, 1.18)
Vertebral fracture	CCAE	1.37 (1.12, 1.67)	1.25 (0.96, 1.62)
	MDCR	1.01 (0.87, 1.17)	0.95 (0.76, 1.18)
	OPTUM-DOD	1.18 (1.04, 1.34)	1.11 (0.98, 1.27)
	OPTUM-EHR	1.19 (1.06, 1.32)	1.14 (0.92, 1.42)
	Meta-analysis	1.17 (1.05, 1.31)	1.10 (1.00, 1.21)

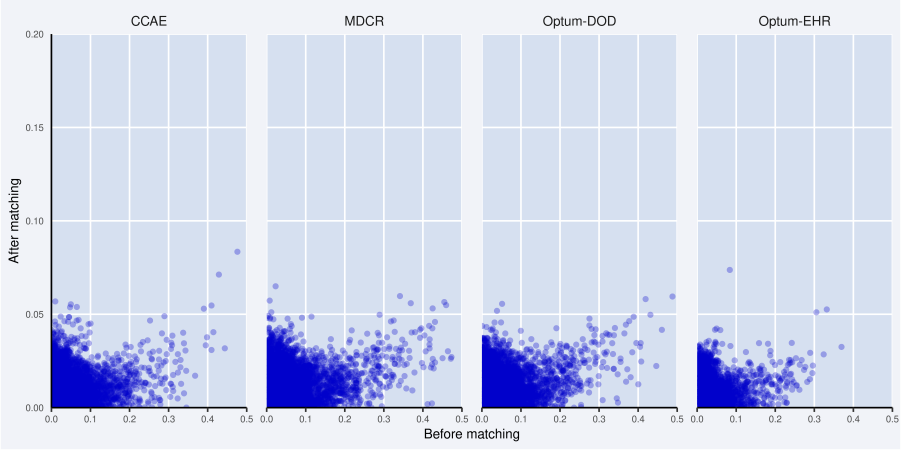

Figure 6.2: test

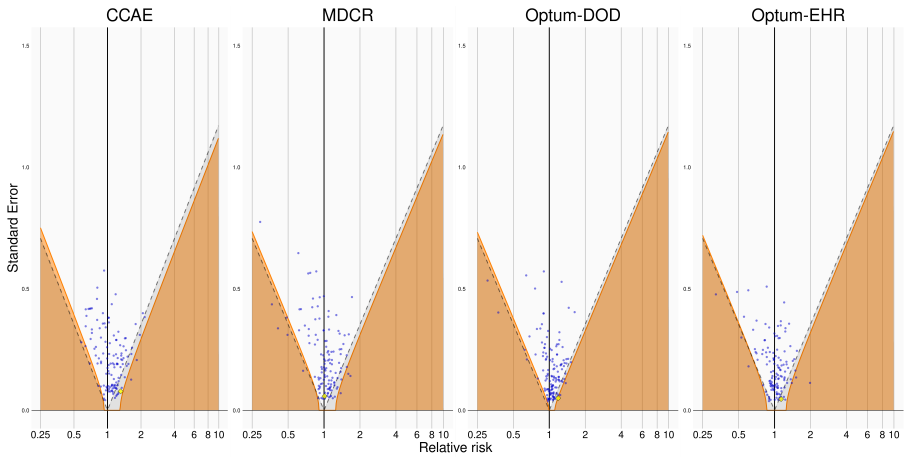

Figure 6.3: test

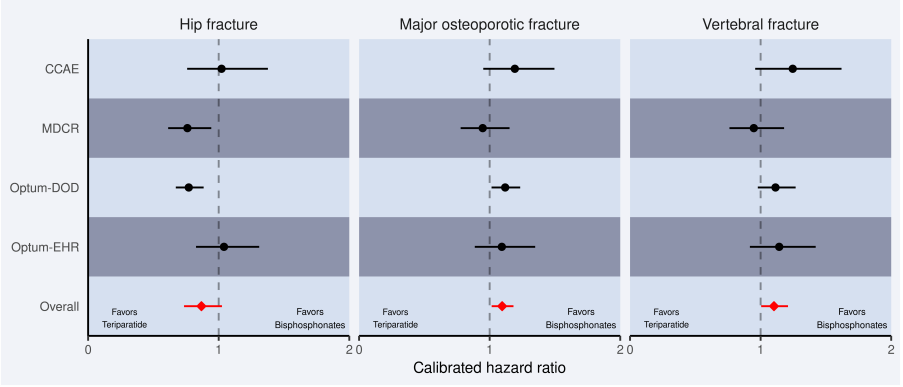

Figure 6.4: test

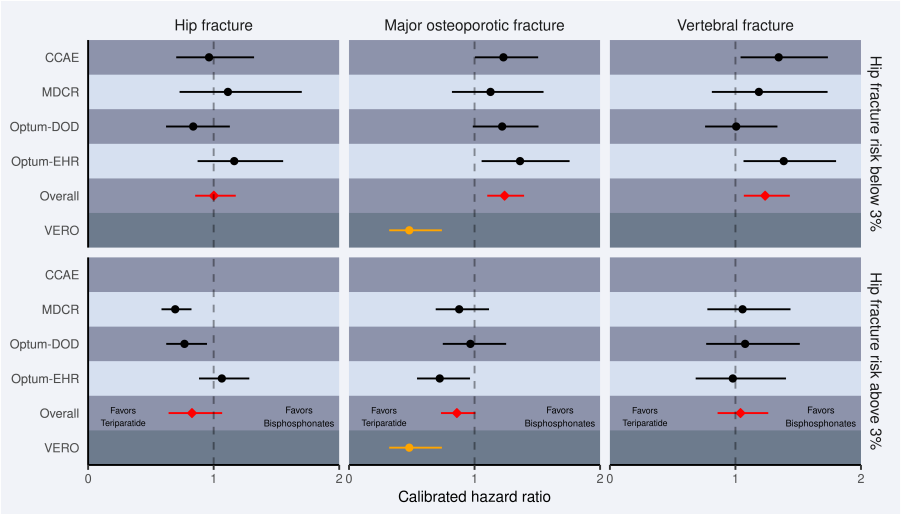

Figure 6.5: test

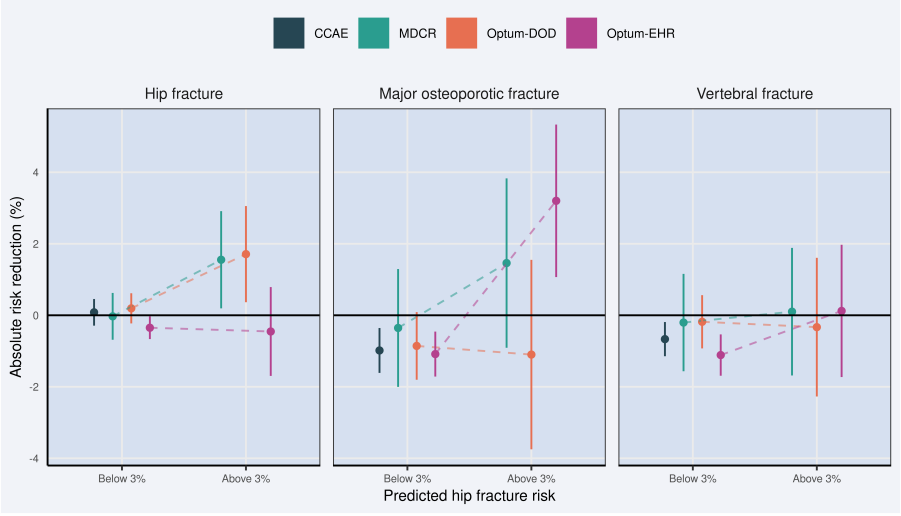

Figure 6.6: test

CHAPTER 7

General discussion

7.1 Main findings

This thesis had three main research aims (Box 7.1). In Chapter 1 we reviewed the available literature on predictive approaches to treatment effect heterogeneity. In Chapter 2 we presented a standardized framework for the assessment of treatment effect heterogeneity using a risk-based approach in observational data. We applied this framework in two different clinical settings. First, we evaluated effect heterogeneity of thiazides or thiazide-like diuretics compared to angiotensin-converting enzyme inhibitors for the treatment of hypertension (Chapter 2). Second, we evaluated treatment effect heterogeneity of teriparatide compared to oral bisphosphonates for patients with osteoporosis (Chapter 6). In Chapter 3 we presented risk-based approaches for predicting individualized treatment effect. Finally, in Chapters 4 and 5 we presented the development and validation of models for prediction of baseline risk in melanoma and COVID-19 patients.

Box 7.1: Results regarding the aims of this thesis

Aim 1: *Systematically review the current literature on predictive approaches to treatment effect heterogeneity.*

- We found 36 papers on regression-based methods. Based on the reference class for patient similarity regarding treatment effect heterogeneity, we categorized existing methods to risk modeling (similarity based on risk factors), treatment effect modeling (similarity based on risk factors and effect modifiers), and optimal treatment regime methods (similarity based solely on effect modifiers).

Aim 2: *Develop scalable and reproducible risk-based predictive approaches to the assessment of treatment effect heterogeneity.*

- We developed a standardized five-step framework for evaluating treatment effect heterogeneity within the observational setting using a risk-based approach.
- We found in simulations that a simple linear interaction of baseline risk with treatment adequately predicted absolute treatment effects for individual patients in many common scenarios.

Aim 3: *Apply risk-based methods to better guide medical decisions*

- We developed and validated a prediction model for sentinel node positive melanoma patients.
- We developed and validated a prediction model for predicting mor-

tality and need for intensive care unit admission in patients who present to the emergency department with suspected COVID-19. We implemented the model in publicly available online applications.
- We evaluated effect heterogeneity of teriparatide treatment compared to treatment with oral bisphosphonates in osteoporosis patients using our five-step framework.

7.2 Review of the literature

In our scoping literature review we categorized the 36 extracted papers with regression-based methods using their definition of the reference class, that is, the subset of observed patient characteristics used to describe patient similarity in terms of expected treatment effects. We identified three different approaches to regression-based evaluation of treatment effect heterogeneity: risk-based methods, treatment effect modeling methods, and optimal treatment regime methods (Box 7.2).

Risk-based approaches define patient similarity based solely on risk factors. They can be further divided into risk stratification approaches that rely on the definition of risk-based subgroups of patients and risk magnification approaches that assume a constant relative treatment effect. The latter can also be used to make personalized benefit predictions. In Chapter 3, the strong assumption of constant relative treatment effects was relaxed, allowing for increasingly linear and non-linear interactions of baseline risk with treatment.

Treatment effect modeling methods focus both on risk factors and treatment effect modifiers to predict personalized treatment effects. They are more intuitive, in the sense that they attempt to account for all dimensions of treatment effect heterogeneity. However, statistical power is an important constraint, as multiple treatment-covariate interaction effects need to be estimated. In the presence of well-documented and clinically supported effect modifiers statistical power may suffice, as only a small pre-defined set of interaction effects will be evaluated[14,16,17]. Penalization methods, shrinking treatment-covariate interactions towards 0, have been shown to improve performance of treatment effect modeling methods. For example, Basu et al developed models for predicting individualized benefits and harms from intensive blood pressure treatment[56]. They found that in the case of predicting adverse events—a setting with limited prior knowledge on effect modification—using elastic net penalization improved performance compared to a backward elimination approach. Similar conclusions were drawn in the simulation study of van Klaveren et al, where LASSO penalization resulted in smaller errors compared to

an unpenalized approach, in the presence of true effect modification[90].

Another approach to the estimation of personalized treatment effects is staging[62,108,181]. Staging approaches are two-stage methods that rely on the calibration of first-stage "working" models with a large set of treatment-covariate interactions. Künzel et al, focusing on machine learning approaches, proposed an organization of staging methods, categorizing them into A-learners and S-learners[62]. Assuming binary treatment assignment, A-learners fit treatment-arm specific models before estimating individualized treatment effects as their difference. S-learners include treatment assignment in the development of the tree-based model. Conditional average treatment effect is then estimated as the difference between setting the treatment indicator to control and active treatment. Finally, they introduced the class of X-learners. First-stage outcome models are fitted separately in each treatment arm to impute coutnerfactual outcomes, thus generating an "observed" treatment effect. Any regular modeling approach can then be used to estimate treatment effects.

Finally, optimal treatment regime methods focus on modeling treatment effect modifiers for the evaluation of treatment effect heterogeneity. Their aim is not to provide personalized treatment effect estimates or to separate patients into subgroups of similar expected treatment effect, but rather to classify them into two categories: patients who benefit from treatment and patients who do not. If there are no major treatment-related harms or costs, they can be used to guide medical decisions. However, in the presence of serious treatment adverse events, these methods may be more challenging to implement, as the effect of baseline risk factors is not taken into account. This means that the baseline risk of the main outcome of interest is not evaluated and, therefore, the absolute risk reduction achieved with treatment cannot be compared to the risk increase for the adverse event in question.

In a similar literature review focusing on subgroup identification, Lipkovich et al divided existing methods into three main categories[182]. The more general approach, global outcome modeling, builds a global model for the outcome of interest often including a large number of covariates and interactions of many of these covariates with treatment. As this approach is quite challenging in its implementation due to low power and limited knowledge on treatment effect modification, Lipkovich et al identified two main simplification approaches. First, global treatment effect modeling approaches focus only on the estimation of treatment-covariate interactions foregoing the estimation of purely prognostic covariate effects. The definition of this set of methods is very similar to the optimal treatment regime category of Chapter 1. Second, local modeling methods focus on the identification of regions of the covariate space with de-

sirable treatment effects. In this way, an outcome model is only required to perform adequately locally, avoiding the need to extend the derived models to covariate regions with limited information. Risk modeling methods presented in Chapter 1 can also be considered as an additional simplification approach, since they assume that treatment effect only interacts with baseline risk.

Our study had several limitations. We focused our review on the clinical trial setting and only regression modeling methods were considered. The analyzed literature was identified through a mix of systematic literature search and suggested literature identified by an expert panel. Though this may have resulted in a more targeted initial selection of reviewed publications, it is possible that not all of the relevant literature was captured. However, the main finding of the review, i.e., a systematic categorization of the predictive methods for the evaluation of treatment effect heterogeneity, is probably robust against missing citations. Furthermore, literature on treatment effect heterogeneity has been growing rapidly since our literature review in 2020[62,91,92,108,183].

Box 7.2: Organization of existing literature

Risk-based methods: *Patient similarity is defined solely based on risk factors.*

- Interactions of baseline risk with treatment are modeled.
- Risk stratification approaches identify risk of homogeneous treatment effect.
- Risk magnification approaches assume a constant relative treatment effect

Treatment effect modeling methods: *Patient similarity is defined based on both risk factors and treatment effect modifiers.*

- The main effects of risk factors are modeled along with interactions of effect modifiers with treatment.
- Statistical power is an important constraint

Optimal treatment regime methods: *Patient similarity is primarily defined based on treatment effect modifiers.*

- Rely on accurately estimating interactions of effect modifiers with treatment.
- Aim to separate patients into two categories: those who benefit from the treatment under study and those who do not.
- In the presence of serious treatment-related harms may be challenging to implement.

7.3 Risk-based predictive approaches to treatment effect heterogeneity

In Chapter 2, we proposed a five-step standardized framework for evaluating treatment effect heterogeneity within the observational setting using a risk-based approach (box 7.3). We also developed a software package for performing such analyses. We made our software compatible with databases mapped to OMOP-CDM to allow for scalability and reproducibility of the analyses. The potential of this approach was demonstrated in a small-scale analysis of effect heterogeneity in the treatment of hypertension, while a more thorough application in the treatment of osteoporosis was presented in Chapter 6.

For the definition of the research aim we need to clearly define the set of characteristics required to qualify patients for any of the following sets: the treatment set (i.e. the patients within a database receiving the treatment under study); the comparator set (i.e., the set of patients within a database receiving the control treatment); one or more outcome sets (i.e., the sets of patients within a database that experience the outcomes of interest). Once the relevant databases to be included in the study have been identified, we can proceed with patient extraction in order to generate the study population for each database.

Consecutively, for each database, we develop prediction models for the outcomes of interest. We develop the prediction models on the propensity score matched subsets of the study populations, in order to avoid overfitting to one of the treatment arms. We then use the predictions of these models to stratify the study population of each database into a predefined number of risk groups.

Within risk strata we develop separate propensity score models. We then estimate relative and absolute treatment effects within strata, adjusting for the observed confounding using the derived risk-stratum specific propensity scores. At this point it is important to analyze diagnostics for the validity of the derived estimates. Adjustment for observed confounding can be evaluated both at database level and within risk strata by plotting the standardized mean differences of the study covariates before and after propensity score adjustment. The presence of unobserved confounding can be evaluated using negative control analyses[27,29,30,103]. These use treatment effect estimates on a large number of causally unrelated outcomes to any of the treatments under study to approximate the observed null distribution and contrast it to the theoretical one. Once all the results have been generated and evaluated for their validity, risk stratified forest plots of treatment effects on both the relative and the absolute scale can be generated in all available databases.

Box 7.3: Standardized framework

Framework steps

1. *Definition of the research aim*
2. *Identification of the relevant databases*
3. *Development of prediction model(s) for the outcome(s) of interest*
4. *Estimation of relative and absolute treatment effects within strata of predicted outcome risk*
5. *Presentation of the results*

Demonstration

1. Compare the effect of thiazide or thiazide-like diuretics to the effect of ACE inhibitors in patients with established hypertension with respect to 12 outcomes (acute myocardial infarction; hospitalization with heart failure; ischemic or hemorrhagic stroke; acute renal failure; kidney disease; cough; hyperkalemia; hypokalemia; gastrointestinal bleeding; hyponatremia; hypotension; angioedema).
2. IBM MarketScan Commercial Claims and Encounters, IBM MarketScan Multi-State Medicaid, and IBM MarketScan Medicare Supplemental Beneficiaries.
3. In each database, develop prediction models for acute myocardial infarction on the 1:1 propensity score-matched subset of the pooled thiazide or thiazide-like diuretics and ACE inhibitor treatment arms.
4. In each database, stratify the population in patients below 1%, between 1% and 1.5% and above 1.5% acute myocardial infarction risk. Within risk groups estimate relative treatment effects from proportional hazards models and absolute treatment effects from the difference of Kaplan-Meier estimates on day 730 after treatment initiation. Adjust for propensity scores for all estimates.
5. Overall treatment effects favoring ACE inhibitors were mainly driven by patients with predicted acute myocardial risk above 1.5%.

The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement aimed to provide guidance on the conduct of predictive analyses of treatment effect heterogeneity[16,17]. Using variation in baseline risk across patient subgroups, the PATH statement presented comprehensive guidance on performing risk-based evaluation of treatment effect heterogeneity. It also identified several knowledge gaps that need to be addressed in future research.

Among many others, the need for further research in the evaluation of treatment effect heterogeneity in observational rather than experimental/randomized data and methods for individualization of predicted treatment effect were highlighted. The proposed framework provides an extension of the PATH statement to the observational setting using a standardized approach. The open-source software we developed for implementing our framework enables highly complex analyses allowing for multiple stratification schemes and analysis approaches. It is also highly scalable as any set of analyses could be implemented across a global network of observational databases mapped to OMOP-CDM.

With the development of the framework presented in Chapter 2 our aim was to enable the evaluation of treatment effect heterogeneity using readily available observational patient data. Despite its great promise, however, observational data suffers from major imbalances in treatment assignment (observed and unobserved confounding), great disparities in the level clinical information is captured between different vendors, and high-dimensionality[184,185]. Consequently, traditional statistical inference methods often fail to produce unbiased treatment effect estimates. Despite great advances in methods research with observational data, strong assumptions—often unverifiable—are required to overcome very fundamental design limitations[39]. These issues may be compounded when evaluating treatment effect heterogeneity, mainly because of the risk of conflating confounding and effect modification.

In the development of the framework in Chapter 2, considerable attention is given to the evaluation of diagnostics at every step of the process. More specifically, the overlap of the propensity score distributions between treatment arms along with the balance in standardized covariate means after propensity score adjustment are evaluated. However, these approaches only focus on adjustment for observed confounding, i.e. confounding that can be eliminated by conditioning on observed covariates. Consequently, both the overall and the risk-stratified treatment effect estimates, can still be biased due to unobserved confounding. Using negative control analyses we can evaluate the presence of unobserved confounding and—to some extent—calibrate the derived estimates.

Risk stratified approaches to treatment effect heterogeneity estimate treatment effects for groups of patients with more similar, but not necessarily equal, baseline risk. Consequently, with larger variability in predicted risk and strong treatment effects, strata-specific estimates of treatment effect may not apply to individuals. In Chapter 3, we compared regression-based methods for personalizing predictions of treatment effect using baseline risk in a simulation study. Focusing on the RCT setting, methods assuming a constant relative treatment effect, or a linear interaction of the linear predictor for risk with

treatment, or finally non-linear interactions in the form of restricted cubic splines were compared. Additionally, an adaptive approach using Akaike's Information Criterion was also evaluated.

No single method outperformed the other methods across all scenarios. However, a linear interaction model performed adequately in most scenarios, while the more flexible restricted cubic splines methods and the adaptive methods required larger sample sizes. The findings of our simulations express the trade-off between the advantages of flexibly modeling the relationship between baseline risk and treatment effect and the disadvantages of overfitting this relationship to the sample at hand.

Although risk is a mathematical determinant of treatment effect, an important limitation of risk-based methods for the evaluation of treatment effect heterogeneity is the strong assumptions on the relationship between treatment effect and risk factors. Since treatment benefit is modeled as a function of predicted baseline risk, all risk factors are assumed be similarly associated with treatment effect. More specifically, the effects of risk factors on treatment effect are assumed to be proportional to their effect on baseline risk[110]. Although models allowing for individual treatment-covariate interactions may be considered more realistic, estimation of these interaction effects can be very challenging[22,186,187]. Randomized controlled trials are often adequately powered for the detection of a main treatment effect and not for the estimation of interaction effects. Consequently, interaction effects are often overestimated, leading to major overestimation of treatment effect heterogeneity ("overfitting").

Therefore, in settings with smaller sample sizes, in the absence of information on treatment effect modification, or in the presence of large number of candidate effect modifiers, the adoption of a risk-based approach for the evaluation of treatment effect heterogeneity is appropriate. In these settings, the bias introduced by—potentially falsely—assuming that treatment effect is a function of baseline risk may still result in more accurate representation of treatment effect heterogeneity, compared to a highly variable treatment effect modeling approach. With larger sample sizes or well-studied treatment effect modification, treatment effect modeling may be the optimal approach. Again, special care, such as penalization of interaction effects may still be required as a measure against overfitting[90]. In general, there is no single approach that can universally outperform all others, but it grossly depends on the setting at hand.

In our simulation study we only considered risk modeling approaches and did not compare them with treatment effect modeling methods. Most of the considered simulation scenarios assumed the presence of true interactions of treat-

ment with baseline risk, without allowing for interactions of treatment with individual predictors. In a small set of additional simulation scenarios, the performance of the considered risk-based methods was evaluated in the presence of true treatment-covariate interactions. Although the conclusions about the optimal risk-based approaches were similar, the errors for all approaches increased considerably. In these scenarios, using a treatment effect modeling approach with appropriate penalization on the interaction effects may have performed better with larger sample size and/or strong treatment-covariate interactions.

7.4 Applications

Box 7.4: Summary of applications

We carried out the following applications of risk-based methods to better guide medical decisions:

- Prediction of 5-year risk of recurrence, distant metastasis, and overall mortality in patients with sentinel node postitive melanoma
- Prediction of 28-day risk of ICU admission and death in patients presenting at the emergency department with suspected COVID-19
- Risk-based evaluation of teriparatide treatment effect heterogeneity compared to treatment with oral bisphosphonates in patients with osteoporosis

We developed and externally validated a nomogram for the prediction of 5-year risk of recurrence, distant metastasis, and overall mortality for positive sentinel node (SN) melanoma patients. We initially fitted a Cox regression model for recurrence using four baseline covariates. Models for distant metastasis and overall mortality were derived by recalibrating the baseline hazard and the slope of the recurrence prediction model. As the MSLT-II trial (**REF**) found no survival benefit for completion lymph node dissection the number of involved non-SNs will often not be available when risk stratifying positive SN melanoma patients. In our prediction model we found minor performance drop when using SN tumor burden as a surrogate. In addition, the developed prediction models were better able to discriminate high from low risk patients compared to the American Joint Committee on Cancer staging system, thus identifying a more robust low-risk group in whom it may be justified to forego adjuvant therapy.

We developed models for the prediction of 28-day mortality and admission

to the intensive care unit (ICU) in patients presenting at the emergency department with suspected COVID-19. The prediction models were developed in patients from four hospitals in the Netherlands during the first COVID-19 wave (March through August 2020) and temporally validated on patients of the second wave (September through December 2020). The proposed models were based on quickly and objectively obtainable predictors (**REFS**). Prediction of ICU admission risk is challenging as it depends on national, regional or hospital practices, as well ICU bed availability. In addition, as was already pointed out in Chapter 5, patients admitted to the ICU tended to be younger than the patients being discharged due to decisions not to admit frail patients. For the development of our prediction model for ICU admission, the strong correlation between death and need for intensive care was exploited, by recalibrating the mortality prediction model for the outcome of ICU admission. The prediction models displayed good performance (discrimination and calibration) in the validation set, were easy to use, and were freely available online and in mobile applications.

The prediction models of Chapters 4 and 5 are applicable to the populations from which the model development samples were drawn. The general problem of transportability of prediction models in space and time is an important issue, that needs to be considered when evaluating prediction models[188,189]. Patient populations can vary substantially over time as they become older, In addition, the characteristics of populations located in different places may be significantly different compared to the model development population – beyond those represented in the prediction model – and disparities in data capture may also be present. To give insight into its transportability, we externally validated the EORTC-DeCOG nomgram of Chapter 4 in data from Germany and adopted a leave-one-center-out validation approach for the COPE prediction model of Chapter 5.

Prediction models for COVID-19 are not easily transportable to settings outside the ones they were actually developed[190]. Response to COVID-19 has been characterized with large geographical and temporal disparities (**REF**), as have the severity and progression of disease. Therefore, our developed prediction model of Chapter 5 could be safely applied to aid medical decision making in the Netherlands during the earlier stages of the pandemic. However, its generalizability to other healthcare systems needs to be explored. In addition, evaluation of its temporal validity in subsequent periods is required and, if found necessary, the model should be updated. Recent work has shown promising results with the application of a model updating framework[191].

7.5 Future research

A review of the fast-growing literature in the observational setting using the suggested categorization approach needs to be carried out. Due to the added complexity (large patient numbers, large number of captured covariates, data quality issues, and many more) focus has shifted from regression modeling methods to more automated machine-learning approaches. Consequently, treatment effect modeling and optimal treatment regime methods have become more prevalent[92,192,193]. The categorization approach suggested can be used to guide this endeavor, while also it can be further generalized to account for the fundamental differences between settings.

A large-scale application of the framework for risk-based assessment of treatment effect heterogeneity in the observational setting should be carried out to better demonstrate its potential for providing better insight to the derived overall treatment effect estimates. The comparison of thiazides or thiazide-like diuretics to ACE inhibitors presented as a demonstration in the presentation of the framework was based on a limited set of treatments and outcomes used in a very large observational study[33]. That study compared first-line treatments for hypertension across an extensive network of observational databases from around the world. Extension of the small-scale application to the entire set of comparisons considered in that study is a realistic aim.

The PATH statement rightly suggested developing the internal prediction model on the pooled study population to avoid biases in the risk stratified treatment effect estimates, both on the relative and the absolute scale[15,18,22]. Despite this being straightforward to implement within the RCT setting, the systematic differences of patient characteristics between treatment arms in the observational setting complicate risk stratification for the evaluation of treatment effect heterogeneity. In our framework we developed the internal prediction model on the propensity score matched subset of the study population. Even though this approach achieves balance between treatment arms, it effectively modifies the target population of the prediction model. Consequently, further research on the modeling approaches of the framework are required.

The proposed approaches to modeling interactions of baseline risk with treatment compared in the simulation study of Chapter 3 need to be extended to the observational setting and evaluated in an extensive simulation study. Extensive literature on machine learning algorithms for estimating conditional average treatment effects and for correcting overfitted regression-based treatment effect modeling approaches provides a broad set of candidate methods for the evaluation of treatment effect heterogeneity on the observational

setting[91,108,109]. A head-to-head comparison of these methods and the proposed risk-based approaches can provide further insights into their relative performance and help guide model selection and implementation in different settings.

References

1. Williams, B. *et al.* 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH) . *European Heart Journal* **39**, 3021–3104 (2018).

2. Kanis, J. A. *et al.* Algorithm for the management of patients at low, high and very high risk of osteoporotic fractures . *Osteoporosis International* **31**, 1–12 (2019).

3. Moons, K. G. M. *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker . *Heart* **98**, 683–690 (2012).

4. Steyerberg, E. W. *Clinical Prediction Models.* (Springer International Publishing, 2019). doi:10.1007/978-3-030-16399-0.

5. Kravitz, R. L., Duan, N. & Braslow, J. Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages . *The Milbank Quarterly* **82**, 661–687 (2004).

6. Greenfield, S., Kravitz, R., Duan, N. & Kaplan, S. H. Heterogeneity of Treatment Effects: Implications for Guidelines, Payment, and Quality Assessment . *The American Journal of Medicine* **120**, S3–s9 (2007).

7. Rothwell, P. M. Can overall results of clinical trials be applied to all patients? *The Lancet* **345**, 1616–1619 (1995).

8. Holland, P. W. Statistics and Causal Inference. *Journal of the American Statistical Association* **81**, 945–960 (1986).

9. Kahneman, D. & Lovallo, D. Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science* **39**, 17–31 (1993).

10. Guyatt, D. L. ;. C., Gordon H.; Sackett. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. . *Jama* **270**, 2598–2601 (1993).

11. Sun, X., Ioannidis, J. P. A., Agoritsas, T., Alba, A. C. & Guyatt, G. How to Use a Subgroup Analysis: Users' Guide to the Medical Literature. *Jama* **311**, 405–411 (2014).

12. Rothwell, P. M. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation . *The Lancet* **365**, 176–186 (2005).

13. Kent, D. M., Rothwell, P. M., Ioannidis, J. P., Altman, D. G. & Hayward, R. A. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal . *Trials* **11**, (2010).

14. Kent, D. M., Steyerberg, E. & Klaveren, D. van. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects . *Bmj* **363**, k4245 (2018).

15. Burke, J. F., Hayward, R. A., Nelson, J. P. & Kent, D. M. Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials . *Circulation: Cardiovascular Quality and Outcomes* **7**, 163–169 (2014).

16. Kent, D. M. *et al.* The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement . *Annals of Internal Medicine* **172**, 35 (2019).

17. Kent, D. M. *et al.* The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration . *Annals of Internal Medicine* **172**, W1 (2019).

18. Abadie, A., Chingos, M. M. & West, M. R. Endogenous Stratification in Randomized Experiments. *The Review of Economics and Statistics* **100**, 567–580 (2018).

19. Califf, R. M. *et al.* Selection of thrombolytic therapy for individual patients: Development of a clinical model . *American Heart Journal* **133**, 630–639 (1997).

20. Dahabreh, I. J., Hayward, R. & Kent, D. M. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence . *International Journal of Epidemiology* **45**, 2184–2193 (2016).

21. Venekamp, R. P., Rovers, M. M., Hoes, A. W. & Knol, M. J. Subgroup analysis in randomized controlled trials appeared to be dependent on whether relative or absolute effect measures were used . *Journal of Clinical Epidemiology* **67**, 410–415 (2014).

22. Klaveren, D. van, Steyerberg, E. W., Serruys, P. W. & Kent, D. M. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects . *Journal of Clinical Epidemiology* **94**, 59–68 (2018).

23. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects . *Biometrika* **70**, 41–55 (1983).

24. D'Agostino Jr., R. B. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group . *Statistics in Medicine* **17**, 2265–2281 (1998).

25. Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies . *Multivariate Behavioral Research* **46**, 399–424 (2011).

26. Dahabreh, I. J. *et al.* Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes . *European Heart Journal* **33**, 1893–1901 (2012).

27. Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A. & Madigan, D. Interpreting observational studies: why empirical calibration is needed to correct p-values . *Statistics in Medicine* **33**, 209–218 (2014).

28. Ryan, P. B. *et al.* Defining a Reference Set to Support Methodological Research in Drug Safety . *Drug Safety* **36**, 33–47 (2013).

29. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Robust empirical calibration of p-values using observational data. *Statistics in Medicine* **35**, 3883–3888 (2016).

30. Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D. & Suchard, M. A. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data . *Proceedings of the National Academy of Sciences* **115**, 2571–2577 (2018).

31. Hripcsak, G. *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers . *Studies in health technology and informatics* **216**, 574 (2015).

32. Matcho, A., Ryan, P., Fife, D. & Reich, C. Fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model . *Drug safety* **37**, 945–959 (2014).

33. Suchard, M. A. *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis . *The Lancet* **394**, 1816–1826 (2019).

34. Schwartz, D. & Lellouch, J. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of chronic diseases* **20**, 637–648 (1967).

35. Ford, J., Ian; Norrie. Pragmatic Trials. *The New England journal of medicine* **375**, 454–463 (2016).

36. Caplan, L. R. Evidence based medicine: concerns of a clinical neurologist. *Journal of neurology, neurosurgery, and psychiatry* **71**, 569–574 (2001).

37. Kent, G. D., David M.; Kitsios. Against pragmatism: on efficacy, effectiveness and the real world. *Trials* **10**, 48–48 (2009).

38. Kent, D. M. & Hayward, R. A. Limitations of Applying Summary Results of Clinical Trials to Individual Patients . *Jama* **298**, 1209 (2007).

39. Varadhan, R., Segal, J. B., Boyd, C. M., Wu, A. W. & Weiss, C. O. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research . *Journal of Clinical Epidemiology* **66**, 818–825 (2013).

40. Daudt, H. M., Mossel, C. van & Scott, S. J. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework . *BMC Medical Research Methodology* **13**, 48 (2013).

41. Rothman, K. J., Greenland, S., Lash, T. L., *et al. Modern epidemiology*. vol. 3 (Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008).

42. Dorresteijn, F. L. J. ;. R., Jannick A. N; Visseren. Estimating treatment effects for individual patients based on the results of randomised clinical trials . *BMJ (Clinical research ed.)* **343**, d5888–d5888 (2011).

43. Vickers, M. W. ;. S., Andrew J.; Kattan. Method for evaluating prediction models that apply the results of randomized trials to individual patients . *Trials* **8**, 14–14 (2007).

44. Julien, J. A., Marilyse; Hanley. Profile-specific survival estimates: making reports of clinical trials more patient-relevant. . *Clinical trials (London, England)* **5**, 107–115 (2008).

45. Hayward, R. A., Kent, D. M., Vijan, S. & Hofer, T. P. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis . *BMC Medical Research Methodology* **6**, (2006).

46. Iwashyna, J. F. ;. S., Theodore J.; Burke. Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care. . *American journal of respiratory and critical care medicine* **192**, 1045–1051 (2015).

47. Kent, D. M. *et al.* Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials . *International Journal of Epidemiology* **45**, dyw118 (2016).

48. Kozminski, J. T. ;. N., Michael A.; Wei. Baseline characteristics predict risk of progression and response to combined medical therapy for benign prostatic hyperplasia (BPH). . *BJU international* **115**, 308–318 (2014).

49. Sussman, J. B., Kent, D. M., Nelson, J. P. & Hayward, R. A. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program . *Bmj* **350**, h454–h454 (2015).

50. Upshaw, M. A. ;. van K., Jenica N.; Konstam. Multistate Model to Predict Heart Failure Hospitalizations and All-Cause Mortality in Outpatients With Heart Failure With Reduced Ejection Fraction: Model Derivation and External Validation. . *Circulation. Heart failure* **9**, Na–na (2016).

51. Groenwold, K. G. M. ;. P., Rolf H. H.; Moons. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings . *Journal of clinical epidemiology* **78**, 90–100 (2016).

52. Follmann, M. A., Dean; Proschan. A multivariate test of interaction for use in clinical trials. *Biometrics* **55**, 1151–1155 (1999).

53. Kovalchik, R. W., Stephanie; Varadhan. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. . *Statistics in medicine* **32**, 4906–4923 (2013).

54. Serruys, M.-C. K., Patrick W.; Morice. Percutaneous Coronary Intervention versus Coronary-Artery Bypass Grafting for Severe Coronary Artery Disease . *The New England journal of medicine* **360**, 961–972 (2009).

55. Klaveren, Y. F. van, David; Vergouwe. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. . *Journal of clinical epidemiology* **68**, 1366–1374 (2015).

56. Basu, J. B. ;. R., Sanjay; Sussman. Benefit and harm of intensive blood pressure treatment: Derivation and validation of risk models using data from the SPRINT and ACCORD trials. . *PLoS medicine* **14**, e1002410–NA (2017).

57. Gerstein, M. B., Hertzel C.; Miller. Effects of intensive glucose lowering in type 2 diabetes. *The New England journal of medicine* **358**, 2545–2559 (2008).

58. Ternès, N., Rotolo, F., Heinze, G. & Michiels, S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces . *Biometrical Journal* **59**, 685–701 (2017).

59. Cai, L. W., Tianxi; Tian. Analysis of randomized comparative clinical trial data for personalized treatment selections . *Biostatistics (Oxford, England)* **12**, 270–282 (2010).

60. Claggett, L. C., Brian; Tian. Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. . *Biostatistics (Oxford, England)* **16**, 60–72 (2014).

61. Zhao, L. C., Lihui; Tian. Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association* **108**, 527–539 (2013).

62. Künzel, J. S. ;. B., Sören R.; Sekhon. Metalearners for estimating heterogeneous treatment effects using machine learning . *Proceedings of the National Academy of Sciences of the United States of America* **116**, 4156–4165 (2019).

63. Weisberg, V. P., Herbert I.; Pontes. Post hoc subgroups in clinical trials: Anathema or analytics?: *Clinical trials (London, England)* **12**, 357–364 (2015).

64. Berger, X. S., James O.; Wang. A Bayesian approach to subgroup identification. *Statistical Methods in Medical Research* **30**, 1465–1483 (2014).

65. Qian, S. A., Min; Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics* **39**, 1180–1210 (2011).

66. Zhang, A. A. ;. L., Baqun; Tsiatis. A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* **68**, 1010–1018 (2012).

67. Taylor, W. F., Jeremy M. G.; Cheng. Reader reaction to "a robust method for estimating optimal treatment regimes" by Zhang et al. (2012). . *Biometrics* **71**, 267–273 (2014).

68. Zhang, A. A. ;. D., Baqun; Tsiatis. Estimating optimal treatment regimes from a classification perspective. *Stat* **1**, 103–114 (2012).

69. Foster, J. M. G. ;. K., Jared C.; Taylor. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data . *Biostatistics (Oxford, England)* **16**, 368–382 (2014).

70. Xu, M. Z., Yaoyao; Yu. Regularized outcome weighted subgroup identification for differential treatment effects. . *Biometrics* **71**, 645–653 (2015).

71. Tian, A. A. ;. G., Lu; Alizadeh. A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates . *Journal of the American Statistical Association* **109**, 1517–1532 (2014).

72. Kraemer, H. C. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach . *Statistics in Medicine* **32**, 1964–1973 (2013).

73. Wallace, E. K., Meredith L.; Frank. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. . *JAMA psychiatry* **70**, 1241–1247 (2013).

74. Niles, A. G. ;. K., Andrea N.; Loerinc. Advancing Personalized Medicine: Application of a Novel Statistical Method to Identify Treatment Moderators in the Coordinated Anxiety Learning and Management Study. . *Behavior therapy* **48**, 490–500 (2017).

75. Gunter, J. M., Lacey; Zhu. Variable Selection for Qualitative Interactions in Personalized Medicine While Controlling the Family-Wise Error Rate . *Journal of biopharmaceutical statistics* **21**, 1063–1078 (2011).

76. Petkova, T. S., Eva; Tarpey. Generated effect modifiers (GEM's) in randomized clinical trials. *Biostatistics (Oxford, England)* **18**, 105–118 (2016).

77. Luedtke, M. J., Alex; van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy . *Annals of statistics* **44**, 713–742 (2016).

78. Laan, A. van der, Mark J.; Luedtke. Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. . *Journal of causal inference* **3**, 61–95 (2015).

79. Chakraborty, B., Laber, E. B. & Zhao, Y.-Q. Inference about the expected performance of a data-driven dynamic treatment regime . *Clinical Trials* **11**, 408–417 (2014).

80. Luedtke, M. J., Alex; van der Laan. Evaluating the impact of treating the optimal subgroup. *Statistical methods in medical research* **26**, 1630–1640 (2017).

81. Robins, A., James M.; Rotnitzky. Discussion of 'Dynamic treatment regimes: Technical challenges and applications' . *Electronic Journal of Statistics* **8**, 1273–1289 (2014).

82. Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. A comparison of methods for model selection when estimating individual treatment effects . (2018) doi:10.48550/arxiv.1804.05146.

83. Chen, W., Ghosh, D., Raghunathan, T. E. & Sargent, D. J. Bayesian Variable Selection with Joint Modeling of Categorical and Survival Outcomes: An Application to Individualizing Chemotherapy Treatment in Advanced Colorectal Cancer . *Biometrics* **65**, 1030–1040 (2009).

84. Janes, M. S. ;. B., Holly; Pepe. Measuring the Performance of Markers for Guiding Treatment Decisions. *Annals of internal medicine* **154**, 253–259 (2011).

85. Janes, M. S. ;. M., Holly; Pepe. The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment . *Journal of the National Cancer Institute* **107**, djv157–NA (2015).

86. Huang, P. B. ;. J., Ying; Gilbert. Assessing treatment-selection markers using a potential outcomes framework. . *Biometrics* **68**, 687–696 (2012).

87. Polley, B. K., Mei Yin C.; Freidlin. Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers . *Journal of the National Cancer Institute* **105**, 1677–1683 (2013).

88. Zhao, D. R., Yingqi; Zeng. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association* **107**, 1106–1118 (2012).

89. Zhao, Y. Q. *et al.* Doubly robust learning for estimating individualized treatment with censored data . *Biometrika* **102**, 151–168 (2014).

90. Klaveren, D. van, Balan, T. A., Steyerberg, E. W. & Kent, D. M. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting . *Journal of Clinical Epidemiology* **114**, 72–83 (2019).

91. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *The Annals of Statistics* **47**, (2019).

92. Powers, J. J., Scott; Qian. Some methods for heterogeneous treatment effect estimation in high dimensions . *Statistics in medicine* **37**, 1767–1787 (2018).

93. Louizos, C. *et al.* Causal effect inference with deep latent-variable models. *Advances in neural information processing systems* **30**, (2017).

94. Navar, M. J. ;. R., Ann Marie; Pencina. Use of Open Access Platforms for Clinical Trial Data. *Jama* **315**, 1283–1284 (2016).

95. Ross, J. S. Clinical research data sharing: what an open science world means for researchers involved in evidence synthesis. . *Systematic reviews* **5**, 159–159 (2016).

96. Ross, J. B., Joseph S.; Waldstreicher. Overview and experience of the YODA Project with clinical trial data sharing after 5 years. . *Scientific data* **5**, 180268–180268 (2018).

97. ROTHWELL, P., MEHTA, Z., HOWARD, S., GUTNIKOV, S. & WARLOW, C. From subgroups to individuals: general principles and the example of carotid endarterectomy . *The Lancet* **365**, 256–265 (2005).

98. Kent, D. M., Alsheikh-Ali, A. & Hayward, R. A. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* **9**, (2008).

99. Thune, J. J. *et al.* Simple Risk Stratification at Admission to Identify Patients With Reduced Mortality From Primary Angioplasty . *Circulation* **112**, 2017–2021 (2005).

100. Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G. & Stang, P. E. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* **19**, 54–60 (2012).

101. Rekkas, A. *et al.* Predictive approaches to heterogeneous treatment effects: a scoping review . *BMC Medical Research Methodology* **20**, (2020).

102. Anglemyer, A., Horvath, H. T. & Bero, L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials . *Cochrane Database of Systematic Reviews* **2014**, (2014).

103. Schuemie, M. J., Ryan, P. B., Hripcsak, G., Madigan, D. & Suchard, M. A. Improving reproducibility by using high-throughput observational studies with empirical calibration . *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**, 20170356 (2018).

104. Whelton, P. K. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines . *Hypertension* **71**, (2018).

105. Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B. & Rijnbeek, P. R. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data . *Journal of the American Medical Informatics Association* **25**, 969–975 (2018).

106. Collins, J. B. ;. A., Gary S.; Reitsma. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement . *Annals of internal medicine* **162**, 55–63 (2015).

107. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration . *Annals of Internal Medicine* **162**, W1 (2015).

108. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353–7360 (2016).

109. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests . *Journal of the American Statistical Association* **113**, 1228–1242 (2018).

110. Hoogland, J. *et al.* A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint . *Statistics in Medicine* **40**, 5961–5981 (2021).

111. Rysavy, M. A. *et al.* Should Vitamin A Injections to Prevent Bronchopulmonary Dysplasia or Death Be Reserved for High-Risk Infants? Reanalysis of the National Institute of Child Health and Human Development Neonatal Research Network Randomized Trial . *The Journal of Pediatrics* **236**, 78–85.e5 (2021).

112. Harrell, F. E., Lee, K. L. & Pollock, B. G. Regression Models in Clinical Studies: Determining Relationships Between Predictors and Response . *JNCI Journal of the National Cancer Institute* **80**, 1198–1202 (1988).

113. Austin, P. C. & Steyerberg, E. W. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine* **38**, 4051–4065 (2019).

114. Steyerberg, E. W., Bossuyt, P. M. M. & Lee, K. L. Clinical trials in acute myocardial infarction: Should we adjust for baseline characteristics? . *American Heart Journal* **139**, 745–751 (2000).

115. Glasziou, P. P. & Irwig, L. M. An evidence based approach to individualising treatment. *Bmj* **311**, 1356–1359 (1995).

116. Lu, M., Sadiq, S., Feaster, D. J. & Ishwaran, H. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods . *Journal of Computational and Graphical Statistics* **27**, 209–219 (2018).

117. Farooq, V. *et al.* Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II . *The Lancet* **381**, 639–650 (2013).

118. Takahashi, K. *et al.* Redevelopment and validation of the SYNTAX score II to individualise decision making between percutaneous and surgical revascularisation in patients with complex coronary artery disease: secondary analysis of the multicentre randomised controlled SYNTAXES trial with external cohort validation . *The Lancet* **396**, 1399–1412 (2020).

119. Gershenwald, R. A. ;. H., Jeffrey E.; Scolyer. Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. . *CA: a cancer journal for clinicians* **67**, 472–492 (2017).

120. Balch, J. E. ;. S., Charles M.; Gershenwald. Final Version of 2009 AJCC Melanoma Staging and Classification. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 6199–6206 (2009).

121. Faries, J. F. ;. C., Mark B.; Thompson. Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma . *The New England journal of medicine* **376**, 2211–2222 (2017).

122. Leiter, R. M., Ulrike; Stadler. Complete lymph node dissection versus no dissection in patients with sentinel lymph node biopsy positive melanoma (DeCOG-SLT): a multicentre, randomised, phase 3 trial . *The Lancet. Oncology* **17**, 757–767 (2016).

123. Leiter, R. M., Ulrike; Stadler. Final analysis of DECOG-SLT trial: Survival outcomes of complete lymph node dissection in melanoma patients with positive sentinel node. . *Journal of Clinical Oncology* **36**, 9501–9501 (2018).

124. Leiter, R. M., Ulrike; Stadler. Final Analysis of DeCOG-SLT Trial: No Survival Benefit for Complete Lymph Node Dissection in Patients With Melanoma With Positive Sentinel Node . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **37**, 3000–3008 (2019).

125. Eggermont, C. U. ;. M., Alexander M. M.; Blank. Adjuvant Pembrolizumab versus Placebo in Resected Stage III Melanoma. *The New England journal of medicine* **378**, 1789–1801 (2018).

126. Long, A. ;. S., Georgina V.; Hauschild. Adjuvant Dabrafenib plus Trametinib in Stage III BRAF-Mutated Melanoma. *The New England journal of medicine* **377**, 1813–1823 (2017).

127. Weber, M. D. V., Jeffrey S.; Mandalà. Adjuvant Nivolumab versus Ipilimumab in Resected Stage III or IV Melanoma. *The New England journal of medicine* **377**, 1824–1835 (2017).

128. Eggermont, V. G., Alexander M. M.; Chiarion-Sileni. Prolonged Survival in Stage III Melanoma with Ipilimumab Adjuvant Therapy. *The New England journal of medicine* **375**, 1845–1855 (2016).

129. Verver, D. *et al.* Risk stratification of sentinel node–positive melanoma patients defines surgical management and adjuvant therapy treatment considerations . *European Journal of Cancer* **96**, 25–33 (2018).

130. Gershenwald, R. H. I. ;. P., Jeffrey E.; Andtbacka. Microscopic Tumor Burden in Sentinel Lymph Nodes Predicts Synchronous Nonsentinel Lymph Node Involvement in Patients With Melanoma . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **26**, 4296–4303 (2008).

131. Lee, R. T.-I., Jonathan H.; Essner. Factors Predictive of Tumor-Positive Nonsentinel Lymph Nodes After Tumor-Positive Sentinel Lymph Node Dissection for Melanoma . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **22**, 3677–3684 (2004).

132. Murali, C. T., Rajmohan; Desilva. Non-Sentinel Node Risk Score (N-SNORE): A Scoring System for Accurately Stratifying Risk of Non-Sentinel Node Positivity in Patients With Cutaneous Melanoma With Positive Sentinel Lymph Nodes . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 4441–4449 (2010).

133. Ploeg, A. C. J. ;. R. van der, Augustinus P. T.; van Akkooi. Prognosis in Patients With Sentinel Node–Positive Melanoma Is Accurately Defined by the Combined Rotterdam Tumor Load and Dewar Topography Criteria . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **29**, 2206–2214 (2011).

134. Ophuis, A. C. J. ;. R., C. M. C. Oude; van Akkooi. Timing of completion lymphadenectomy after positive sentinel node biopsy in patients with melanoma . *The British journal of surgery* **104**, 726–733 (2017).

135. Ophuis, C. R., C. M. C. Oude; Verhoef. The interval between primary melanoma excision and sentinel node biopsy is not associated with survival in sentinel node positive patients – An EORTC Melanoma Group study . *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* **42**, 1906–1913 (2016).

136. Kattan, K. R. ;. A., Michael W.; Hess. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. . *CA: a cancer journal for clinicians* **66**, 370–374 (2016).

137. Buuren, K. van, Stef; Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1–67 (2011).

138. Steyerberg, E. W. *Clinical Prediction Models*. vol. Na Na–na (2009).

139. Harrell, K. L. ;. M., Frank E.; Lee. Multivariable Prognostic Models: Issues In Developing Models, Evaluating Assumptions And Adequacy, And Measuring And Reducing Errors . *Statistics in medicine* **15**, 361–387 (1996).

140. Steyerberg, F. E., Ewout W.; Harrell. Prediction models need appropriate internal, internal-external, and external validation . *Journal of clinical epidemiology* **69**, 245–247 (2015).

141. Akkooi, Z. V. van, Alexander C. J.; Nowecki. Sentinel node tumor burden according to the Rotterdam criteria is the most important prognostic factor for survival in melanoma patients: a multicenter study in 388 patients with positive sentinel nodes. . *Annals of surgery* **248**, 949–955 (2008).

142. Yokota, K. *et al.* Adjuvant therapy with nivolumab versus ipilimumab after complete resection of stage III/IV melanoma: Japanese subgroup analysis from the phase 3 CheckMate 238 study . *The Journal of Dermatology* **46**, 1197–1201 (2019).

143. Hauschild, R. S., Axel; Dummer. Longer Follow-Up Confirms Relapse-Free Survival Benefit With Adjuvant Dabrafenib Plus Trametinib in Patients With Resected BRAF V600–Mutant Stage III Melanoma . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **36**, 3441–3449 (2018).

144. Satzger, U. G., Imke; Leiter. Melanoma-specific survival in patients with positive sentinel lymph nodes: Relevance of sentinel tumor burden . *European journal of cancer (Oxford, England : 1990)* **123**, 83–91 (2019).

145. Kretschmer, H. T., Lutz; Starz. Age as a key factor influencing metastasizing patterns and disease-specific survival after sentinel lymph node biopsy for cutaneous melanoma . *International journal of cancer* **129**, 1435–1442 (2011).

146. Eggermont, S. T., Alexander M. M.; Suciu. Long-Term Results of the Randomized Phase III Trial EORTC 18991 of Adjuvant Therapy With Pegylated Interferon Alfa-2b Versus Observation in Resected Stage III Melanoma . *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **30**, 3810–3818 (2012).

147. Eggermont, A. M. M. *et al.* Long term follow up of the EORTC 18952 trial of adjuvant therapy in resected stage IIB–III cutaneous melanoma patients comparing intermediate doses of interferon-alpha-2b (IFN) with observation: Ulceration of primary is key determinant for IFN-sensitivity . *European Journal of Cancer* **55**, 111–121 (2016).

148. Nishino, N. H. ;. H., Mizuki; Ramaiya. Monitoring immune-checkpoint blockade: response evaluation and biomarker development . *Nature reviews. Clinical oncology* **14**, 655–668 (2017).

149. Roulin, M. B., Didier; Matter. Prognostic value of sentinel node biopsy in 327 prospective melanoma patients from a single institution. . *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* **34**, 673–679 (2007).

150. Richtig, E. N., G.; Richtig. Does the time interval between sentinel lymph node biopsy and completion lymph node dissection affect outcome in malignant melanoma? A retrospective cohort study. . *International journal of surgery (London, England)* **75**, 160–164 (2020).

151. Klemen, G. L., Nicholas D.; Han. Completion lymphadenectomy for a positive sentinel node biopsy in melanoma patients is not associated with a survival benefit. . *Journal of surgical oncology* **119**, 1053–1059 (2019).

152. Brink, J. F., Anniek; Alsma. Prediction models for mortality in adult patients visiting the Emergency Department: a systematic review. . *Acute medicine* **18**, 171–183 (2019).

153. Wynants, B. C., Laure; Van Calster. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal . *BMJ (Clinical research ed.)* **369**, m1328–NA (2020).

154. Sperrin, M., Grant, S. W. & Peek, N. Prediction models for diagnosis and prognosis in Covid-19. *Bmj* **369**, (2020).

155. Moons, R. R., Karel G. M.; Wolff. PROBAST : A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration . *Annals of internal medicine* **170**, W1–w33 (2019).

156. Knight, A. P., Stephen R; Ho. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: development and validation of the 4C Mortality Score. . *BMJ (Clinical research ed.)* **370**, m4334–m4334 (2020).

157. Docherty, E. M. G., Annemarie B; Harrison. Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. . *BMJ (Clinical research ed.)* **369**, m1985–NA (2020).

158. Liang, H. O., Wenhua; Liang. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. . *JAMA internal medicine* **180**, 1081–1089 (2020).

159. King, J. S. R., Joseph T.; Yoon. Development and validation of a 30-day mortality index based on pre-existing medical administrative data from 13,323 COVID-19 patients: The Veterans Health Administration COVID-19 (VACO) Index. . *PloS one* **15**, e0241825–NA (2020).

160. Team, R. C. The R project for statistical computing. *R.*

161. Stone, C. J. & Koo, C.-Y. Additive splines in statistics. *Proceedings of the American Statistical Association Original pagination is p* **45**, 48 (1985).

162. Harrell, F. E. *Regression Modeling Strategies - Regression Modeling Strategies.* vol. Na Na–na (2015).

163. Harrell Jr, F. E. *Rms: Regression modeling strategies.* (2023).

164. Klaveren, M. S. van, David; Gonen. A new concordance measure for risk prediction models in external validation settings. . *Statistics in medicine* **35**, 4136–4152 (2016).

165. Wright, N. C., Saag, K. G., Dawson-Hughes, B., Khosla, S. & Siris, E. S. The impact of the new national bone health alliance (NBHA) diagnostic criteria on the prevalence of osteoporosis in the USA. *Osteoporos Int* **28**, 1225–1232 (2016).

166. Borgström, F. *et al.* Fragility fractures in europe: Burden, management and opportunities. *Arch. Osteoporos.* **15**, 59 (2020).

167. Neer, R. M. *et al.* Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis. *N. Engl. J. Med.* **344**, 1434–1441 (2001).

168. Kendler, D. L. *et al.* Effects of teriparatide and risedronate on new fractures in post-menopausal women with severe osteoporosis (VERO): A multicentre, double-blind, double-dummy, randomised controlled trial. *Lancet* **391**, 230–240 (2018).

169. Díez-Pérez, A. *et al.* Effects of teriparatide on hip and upper limb fractures in patients with osteoporosis: A systematic review and meta-analysis. *Bone* **120**, 1–8 (2019).

170. Reyes, C. *et al.* One and two-year persistence with different anti-osteoporosis medications: A retrospective cohort study. *Osteoporos. Int.* **28**, 2997–3004 (2017).

171. Tian, Y., Schuemie, M. J. & Suchard, M. A. Evaluating large-scale propensity score performance through real-world and synthetic data experiments . *International Journal of Epidemiology* **47**, 2005–2014 (2018).

172. Walker, A. M. *et al.* A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research* **3**, 11–20 (2013).

173. Nguyen, T.-L. *et al.* Double-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Med. Res. Methodol.* **17**, 78 (2017).

174. Normand, S. T. *et al.* Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J. Clin. Epidemiol.* **54**, 387–398 (2001).

175. Lipsitch, M., Tchetgen Tchetgen, E. & Cohen, T. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology* **21**, 383–388 (2010).

176. Rekkas, A. *et al.* A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases. *NPJ Digit. Med.* **6**, 58 (2023).

177. Burn, E. *et al.* Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. *Nat. Commun.* **11**, 5009 (2020).

178. McGrath, L. J. *et al.* Using negative control outcomes to assess the comparability of treatment groups among women with osteoporosis in the united states. *Pharmacoepidemiol. Drug Saf.* **29**, 854–863 (2020).

179. Burge, R. T., Disch, D. P., Gelwicks, S., Zhang, X. & Krege, J. H. Hip and other fragility fracture incidence in real-world teriparatide-treated patients in the united states. *Osteoporos. Int.* **28**, 799–809 (2017).

180. Silverman, S. *et al.* Reduction of hip and other fractures in patients receiving teriparatide in real-world clinical practice: Integrated analysis of four prospective observational studies. *Calcif. Tissue Int.* **104**, 193–200 (2019).

181. Foster, J. C., Taylor, J. M. G. & Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867–2880 (2011).

182. Lipkovich, I., Dmitrienko, A. & B. D'Agostino Sr., R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials . *Statistics in Medicine* **36**, 136–196 (2017).

183. Fan, Q., Hsu, Y.-C., Lieli, R. P. & Zhang, Y. Estimation of Conditional Average Treatment Effects With High-Dimensional Data . *Journal of Business & Economic Statistics* **40**, 313–327 (2022).

184. DeRouen, T. A. Promises and Pitfalls in the Use of 'Big Data' for Clinical Research. *Journal of Dental Research* **94**, 107s–109s (2015).

185. Shah, N. D., Steyerberg, E. W. & Kent, D. M. Big Data and Predictive Analytics: Recalibrating Expectations. *Jama* **320**, 27–28 (2018).

186. Brookes, S. T. *et al.* Subgroup analyses in randomized trials: risks of subgroup-specific analyses;: power and sample size for the interaction test . *Journal of Clinical Epidemiology* **57**, 229–236 (2004).

187. Riley, R. D. *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* **39**, 2115–2137 (2020).

188. Van Calster, B., Steyerberg, E. W., Wynants, L. & Smeden, M. van. There is no such thing as a validated prediction model. *BMC Medicine* **21**, 70 (2023).

189. Luijken, K., Groenwold, R. H. H., Van Calster, B., Steyerberg, E. W. & Smeden, M. van. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective . *Statistics in Medicine* **38**, 3444–3459 (2019).

190. Klaveren, D. van *et al.* Prognostic models for COVID-19 needed updating to warrant transportability over time and space . *BMC Medicine* **20**, 456 (2022).

191. Levy, T. J. *et al.* Development and validation of self-monitoring auto-updating prognostic models of survival for hospitalized COVID-19 patients . *Nature Communications* **13**, 6812 (2022).

192. Tsiatis, A. A., Davidian, M., Holloway, S. T. & Laber, E. B. *Dynamic treatment regimes: Statistical methods for precision medicine*. (CRC press, 2019).

193. Wendling, T. *et al.* Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases . *Statistics in Medicine* **37**, 3309–3324 (2018).

194. Harrell, F. E. *Regression modeling strategies*. *Bios* vol. 330 14 (Springer, 2017).

APPENDIX A

A standardized framework for risk-based assessment of treatment effect heterogeneity

	CCAE			MDCD			MDCR		
Characteristic	THZ	ACE	Std. diff	THZ	ACE	Std. diff	THZ	ACE	Std. diff
Age in years	48.7	48.7	0.00	44.6	44.4	0.01	74.3	74.1	0.03
Sex: female	44.5%	0.442	0.01	58.6%	58.5%	0.00	55.1%	55.6%	-0.01
Medical history (General)									
Chronic obstructive lung disease	1.5%	1.4%	0.01	10.1%	9.4%	0.03	6.9%	6.7%	0.01
Crohn's disease	0.2%	0.2%	0.00	0.3%	0.3%	0.00	0.2%	0.2%	0.00
Depressive disorder	3.6%	3.5%	0.01	10.9%	10.6%	0.01	2.9%	2.9%	0.00
Diabetes mellitus	0.1%	0.2%	-0.03	0.2%	0.4%	-0.04	0.0%	0.0%	-0.01
Gastroesophageal reflux disease	5.5%	5.3%	0.01	9.4%	9.1%	0.01	7.7%	7.4%	0.01
Gastrointestinal hemorrhage	0.4%	0.4%	0.00	0.9%	0.9%	0.00	1.2%	1.2%	0.00
Human immunodeficiency virus infection	0.2%	0.2%	0.00	1.1%	1.1%	0.00	0.0%	0.1%	-0.01
Hyperlipidemia	21.1%	21.5%	-0.01	20.6%	20.8%	-0.01	27.0%	26.8%	0.00
Obesity	8.0%	7.9%	0.00	17.8%	17.6%	0.01	2.9%	2.8%	0.00
Osteoarthritis	1.9%	1.8%	0.01	6.1%	5.8%	0.01	6.2%	5.9%	0.01
Pneumonia	0.4%	0.3%	0.01	1.7%	1.6%	0.00	0.3%	0.3%	0.00
Psoriasis	0.9%	0.8%	0.00	0.8%	0.7%	0.01	0.8%	0.9%	-0.01
Rheumatoid arthritis	0.8%	0.7%	0.01	1.3%	1.3%	0.00	1.5%	1.5%	0.00
Ulcerative colitis	0.2%	0.2%	0.00	0.2%	0.2%	0.00	0.3%	0.3%	0.00
Urinary tract infectious disease	5.3%	5.0%	0.01	10.5%	10.2%	0.01	8.3%	8.0%	0.01
Viral hepatitis C	0.1%	0.1%	0.00	1.8%	1.8%	0.00	0.1%	0.1%	0.00
Medical history (Cardiovascular disease)									
Atrial fibrillation	0.3%	0.3%	0.00	0.9%	0.8%	0.01	3.1%	2.9%	0.02
Cerebrovascular disease	0.4%	0.4%	0.00	0.9%	0.9%	0.00	2.4%	2.4%	0.00
Coronary arteriosclerosis	0.8%	0.9%	-0.01	1.4%	1.3%	0.01	4.6%	4.4%	0.01
Heart disease	0.2%	0.2%	0.00	0.6%	0.6%	0.00	0.6%	0.6%	0.00
Heart failure	0.1%	0.1%	0.00	0.8%	0.7%	0.01	0.2%	0.2%	0.00
Ischemic heart disease	0.0%	0.0%	0.00	0.0%	0.0%	0.00	0.1%	0.1%	-0.01
Pulmonary embolism	0.1%	0.1%	0.00	0.3%	0.3%	0.01	0.1%	0.1%	0.00
Venous thrombosis	0.1%	0.1%	0.00	0.2%	0.2%	0.01	0.3%	0.3%	0.00

Table A.1: Baseline characteristics of patients in CCAE, MDCD, and MDCR after stratification on the propensity scores. THZ:

		Thiazide or thiazide-like diuretics			ACE inhibitors		
Database	Risk group	Patients	Person years	Outcomes	Patients	Person years	Outcomes
CCAE	RG-1	347,892	200,792	368	874,820	550,857	1,500
	RG-2	5,576	2,760	23	37,950	23,408	169
	RG-3	2,358	1,042	14	17,599	9,902	144
MDCD	RG-1	39,144	14,584	22	61,229	28,408	109
	RG-2	7,798	3,371	13	19,066	9,823	78
	RG-3	7,893	3,484	41	26,197	13,250	253
MDCR	RG-1	9,635	6,861	19	22,407	16,157	55
	RG-2	14,944	9,985	48	40,296	29,365	216
	RG-3	13,303	7,796	94	43,149	29,468	461

Table A.2: Number of patients, person years, and events in risk strata of acute MI across all databases. RG-1 represents patients at acute MI risk below 1%, RG-2 represents patients at acute Mi risk between 1% and 1.5%, and RG-3 represents patients at acute MI risk larger than 1.5%.

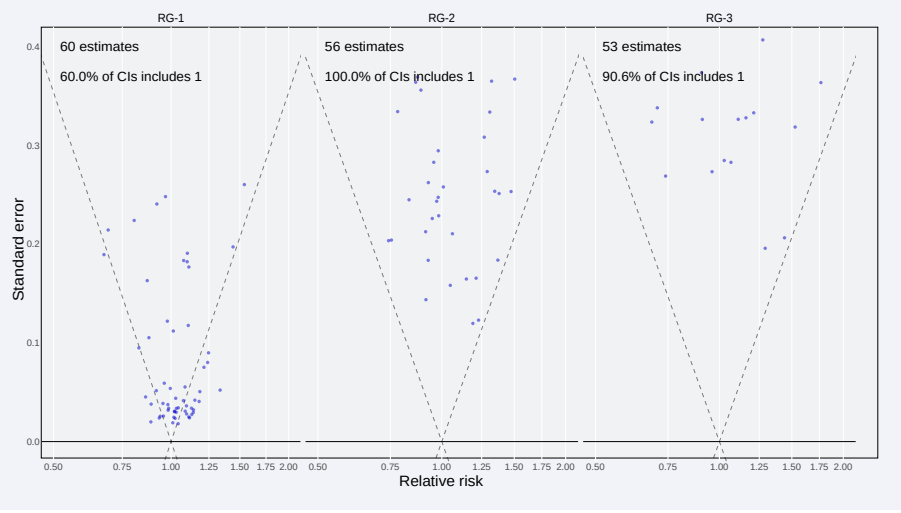

Figure A.1: Systematic error in risk groups of CCAE. Effect size estimates for the negative controls (true hazard ratio = 1) within strata of predicted acute MI risk. Estimates below the diagonal dashed lines are statistically significant (different from the true effect size, alpha = 0.05). RG-1 represents patients in CCAE whose acute MI predicted risk is below 1%; RG-2 represents patients whose acute MI predicted risk is between 1% and 1.5%; RG-3 represents patients whose acute MI predicted risk is larger than 1.5%. A well-calibrated estimator should have the true effect size (HR = 1) within the 95 percent confidence interval 95 percent of times.

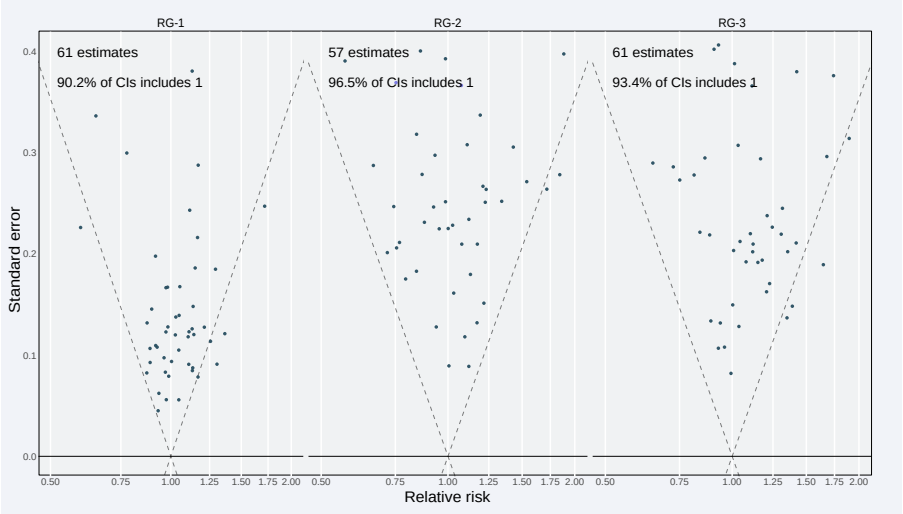

Figure A.2: Systematic error in risk groups of MDCD. Effect size estimates for the negative controls (true hazard ratio = 1) within strata of predicted acute MI risk in MDCD. Estimates below the diagonal dashed lines are statistically significant (different from the true effect size, alpha = 0.05). RG-1 represents patients in MDCD whose acute MI predicted risk is below 1%; RG-2 represents patients whose acute MI predicted risk is between 1% and 1.5%; RG-3 represents patients whose acute MI predicted risk is larger than 1.5%. A well-calibrated estimator should have the true effect size within the 95 percent confidence interval 95 percent of times.

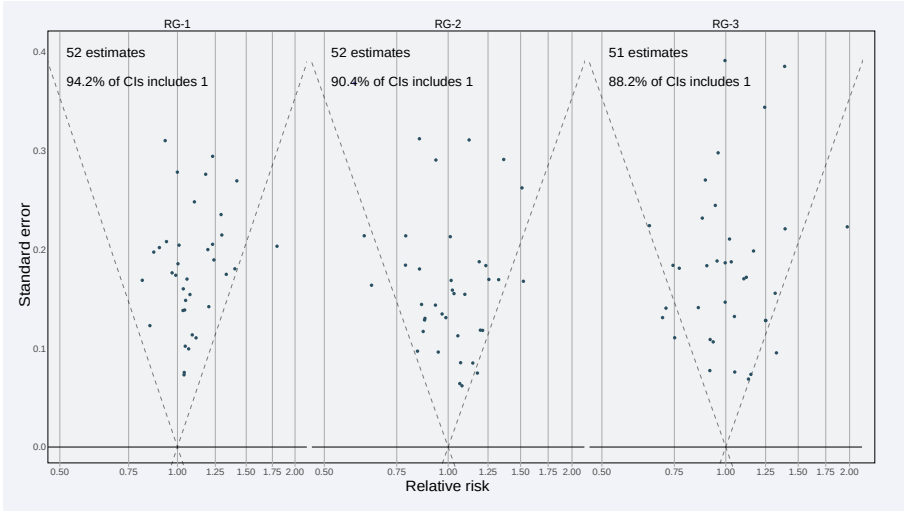

Figure A.3: Systematic error in risk groups of MDCR. Effect size estimates for the negative controls (true hazard ratio = 1) within strata of predicted acute MI risk in MDCD. Estimates below the diagonal dashed lines are statistically significant (different from the true effect size, alpha = 0.05). RG-1 represents patients in MDCD whose acute MI predicted risk is below 1%; RG-2 represents patients whose acute MI predicted risk is between 1% and 1.5%; RG-3 represents patients whose acute MI predicted risk is larger than 1.5%. A well-calibrated estimator should have the true effect size within the 95 percent confidence interval 95 percent of times.

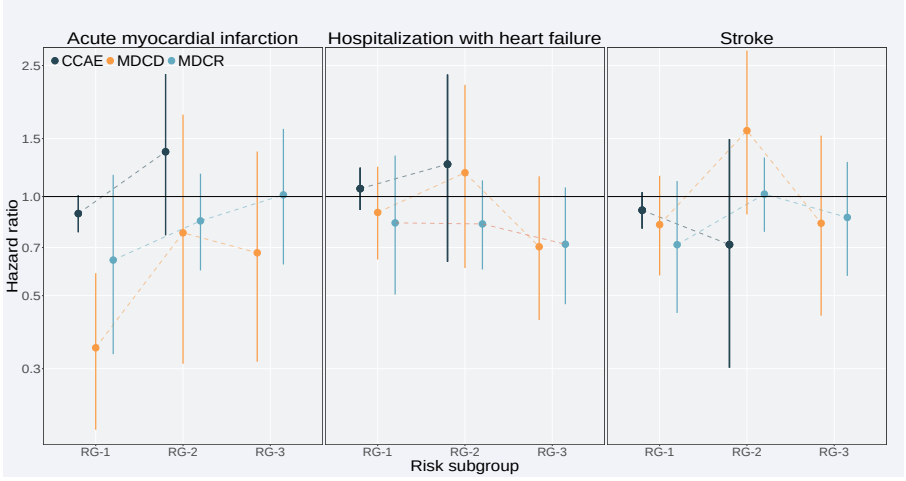

Figure A.4: Relative treatment effects for the main outcomes in patients without cardiovascular disease. Treatment effect heterogeneity in the subset of patients without cardiovascular disease for acute myocardial infarction, hospitalization with heart failure, and stroke on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents %>% the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Hazard ratios estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors.

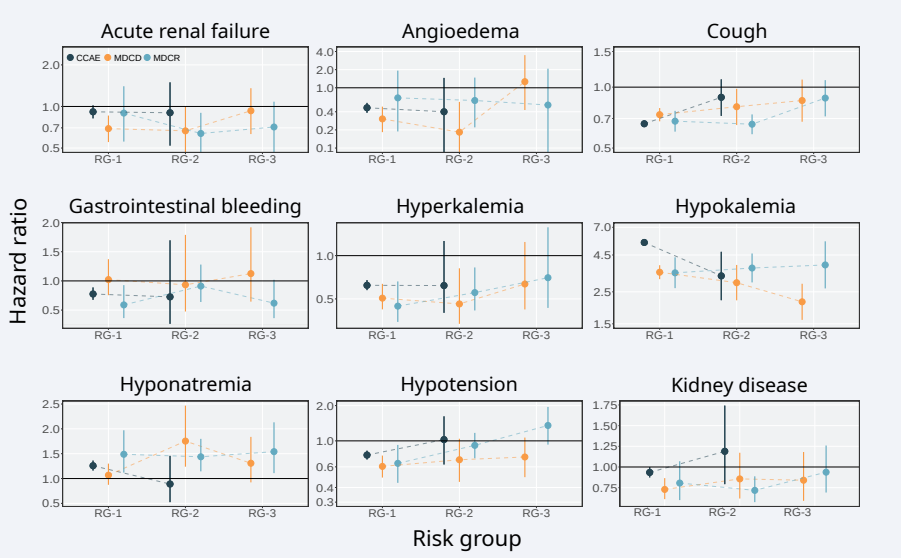

Figure A.5: Relative treatment effects for the safety outcomes in patients without cardiovascular disease. Treatment effect heterogeneity in the subset of patients without cardiovascular disease for acute renal failure, angioedema, cough, gastrointestinal bleeding, hyperkalemia, hypokalemia, hyponatremia, hypotension, and kidney disease on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Hazard ratios estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors.

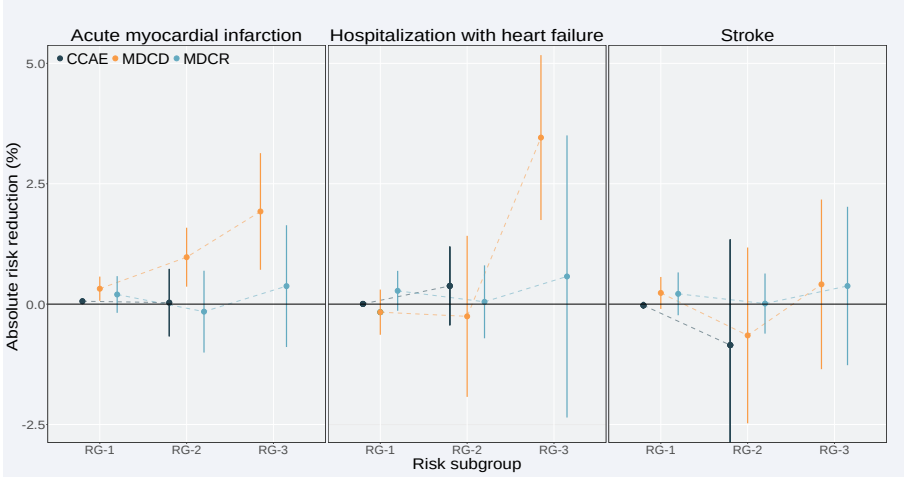

Figure A.6: Absolute treatment effects for the main outcomes in patients without cardiovascular disease. Treatment effect heterogeneity in the subset of patients without cardiovascular disease acute myocardial infarction, hospitalization with heart failure, and stroke on the absolute scale of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Absolute treatment effects estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors.

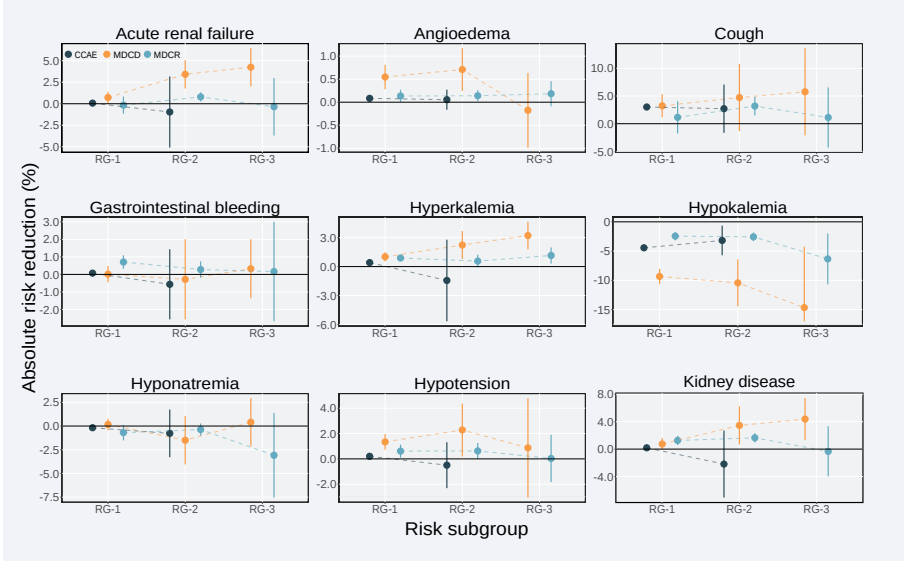

Figure A.7: Absolute treatment effects for the safety outcomes in patients without cardiovascular disease. Treatment effect heterogeneity in the subset of patients without cardiovascular disease for acute renal failure, angioedema, cough, gastrointestinal bleeding, hyperkalemia, hypokalemia, hyponatremia, hypotension, and kidney disease on the absolute scale of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Absolute treatment effects estimated in CCAE, MDCD, and MDCR are represented by blue, green, and orange circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors.

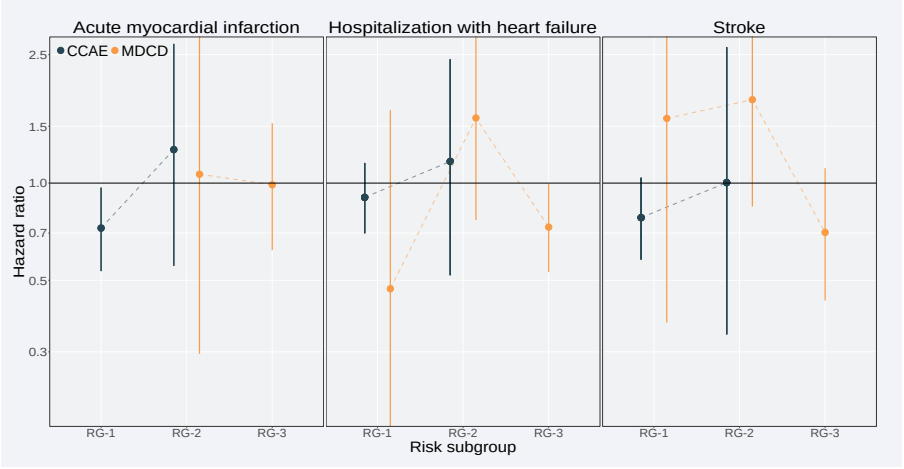

Figure A.8: Relative treatment effects for the main outcomes in patients with cardiovascular disease. Treatment effect heterogeneity in the subset of patients with cardiovascular disease for acute myocardial infarction, hospitalization with heart failure, and stroke on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the subgroup of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors. Hazard ratios estimated in CCAE and MDCD are represented by blue and green circles, respectively. The bars represent 95% confidence intervals. Results in MDCR are not presented because the majority of the patients were at risk above 1.5% for acute MI.

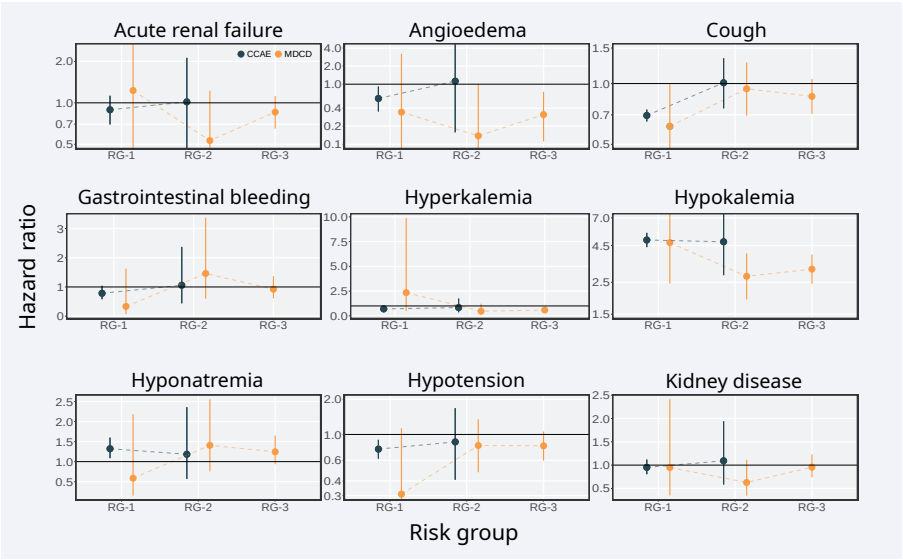

Figure A.9: Relative treatment effects for the safety outcomes in patients with cardiovascular disease. Treatment effect heterogeneity in the subset of patients with cardiovascular disease for acute renal failure, angioedema, cough, gastrointestinal bleeding, hyperkalemia, hypokalemia, hyponatremia, hypotension, and kidney disease on the relative scale (hazard ratios) of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Hazard ratios estimated in CCAE and MDCD are represented by blue and green circles, respectively. The bars represent 95% confidence intervals. Values below 1 favor thiazide or thiazide-like diuretics, while values above 1 favor ACE inhibitors. Results in MDCR are not presented because the majority of the patients were at risk above 1.5% for acute MI.

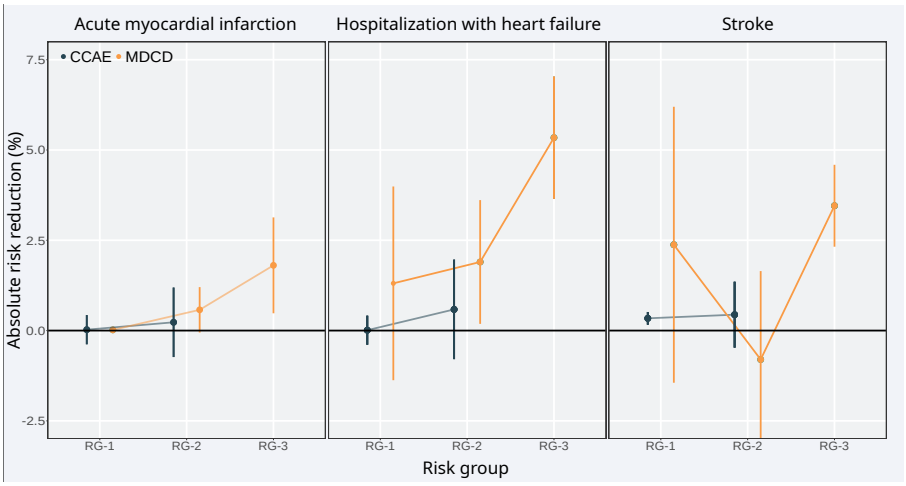

Figure A.10: Absolute treatment effects for the main outcomes in patients with cardiovascular disease. Treatment effect heterogeneity in the subset of patients with cardiovascular disease for the main outcomes of interest on the absolute scale of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1% RG-2 represents the group of patients with acute MI risk between 1% and 1.5% RG-3 represents the group of patients with acute MI risk larger than 1.5% Absolute treatment effects estimated in CCAE and MDCD are represented by blue and green circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors. Results in MDCR are not presented because the majority of the patients were at risk above 1.5% for acute MI.

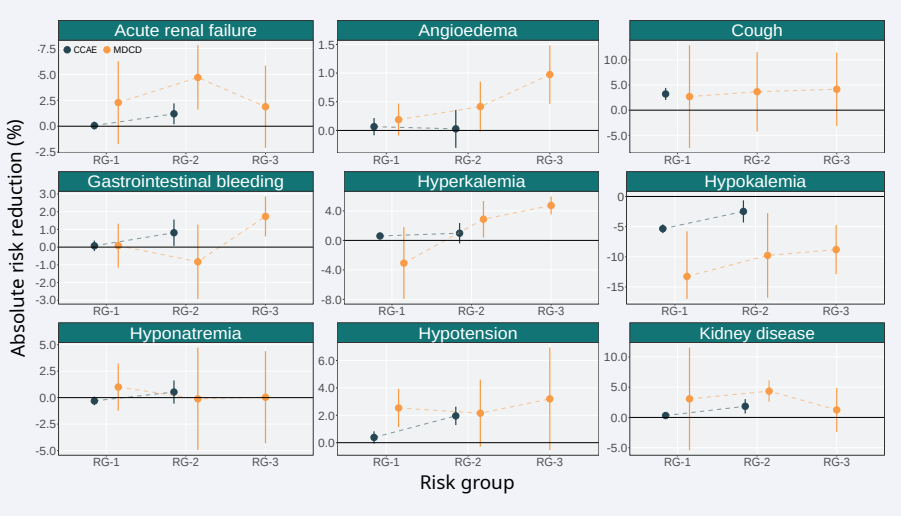

Figure A.11: . Absolute treatment effects for the safety outcomes in patients with cardiovascular disease. Treatment effect heterogeneity in the subset of patients with cardiovascular disease for acute renal failure, angioedema, cough, gastrointestinal bleeding, hyperkalemia, hypokalemia, hyponatremia, hypotension, and kidney disease on the absolute scale of thiazide or thiazide-like diuretics within strata of predicted acute MI risk. RG-1 represents the group of patients with acute MI risk below 1%; RG-2 represents the group of patients with acute MI risk between 1% and 1.5%; RG-3 represents the group of patients with acute MI risk larger than 1.5%. Absolute treatment effects estimated in CCAE and MDCD are represented by blue and green circles, respectively. The bars represent 95% confidence intervals. Values above 0 favor thiazide or thiazide-like diuretics, while values below 0 favor ACE inhibitors. Results in MDCR are not presented because the majority of the patients were at risk above 1.5% for acute MI.

APPENDIX B

Estimating individualized treatment effects from randomized controlled trials

Notation

We observe RCT data (Z, X, Y), where for each patient $Z_i = 0, 1$ is the treatment status, $Y_i = 0, 1$ is the observed outcome and X_i is a set of covariates measured. Let $\{Y_i(z), z = 0, 1\}$ denote the unobservable potential outcomes. We observe $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. We are interested in predicting the conditional average treatment effect (CATE),

$$\tau(x) = E\{Y(0) - Y(1)|X = x\}$$

Assuming that (Z, X, Y) is a random sample from the target population and that $(Y(0), Y(1)) \perp\!\!\!\perp Z|X$, as we are in the RCT setting, we can predict CATE from

$$\begin{aligned}\tau(x) &= E\{Y(0) \mid X = x\} - E\{Y(1) \mid X = x\} \\ &= E\{Y \mid X = x, Z = 0\} - E\{Y \mid X = x, Z = 1\}\end{aligned}$$

Based on an estimate of baseline risk

$$E\{Y \mid X = x, Z = 0\} = g(\widehat{lp}(x))$$

with $\hat{u} = \widehat{lp}(x) = x^t\hat{\beta}$ the linear predictor and g the link function, we predict CATE from

$$\hat{\tau}(x) = g(f(\hat{u}, 0)) - g(f(\hat{u}, 1))$$

where $f(u, z)$ describes interactions of the baseline risk linear predictor with treatment.

Simulation settings

For all patients we observe covariates $X_1, \ldots, X_8$, of which 4 are continuous and 4 are binary. More specifically,

$$\begin{aligned}X_1, \ldots, X_4 &\sim N(0, 1) \\ X_5, \ldots, X_8 &\sim B(1, 0.2)\end{aligned}$$

We first, generate the binary outcomes Y for the untreated patients $(Z = 0)$, based on

$$P(Y(0) = 1 \mid X = x) = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8) = g(lp_0), (\#eq : p0) \quad \text{(B.1)}$$

where

$$g(x) = \frac{e^x}{1 + e^x}$$

For treated patients, outcomes are generated from:

$$P(Y = 1 \mid X = x, Z = 1) = g(lp_1)(\#eq : p1) \tag{B.2}$$

where

$$lp_1 = \gamma_2(lp_0 - c)^2 + \gamma_1(lp_0 - c) + \gamma_0$$

Base-case scenario

The base-case scenario assumes a constant odds ratio of 0.8 in favor of treatment. The simulated datasets are of size $n = 4250$, where treatment is allocated at random using a 50/50 split (80% power for the detection of an unadjusted OR of 0.8, assuming an event rate of 20% in the untreated arm). Outcome incidence in the untreated population is set at 20%. For the development of the prediction model we use the model defined in @ref(eq:p0) including a constant treatment effect. When doing predictions, Z is set to 0. The value of the true β is such that the above prediction model has an AUC of 0.75.

The previously defined targets are achieved when $\beta = (-2.08, 0.49, ..., 0.49)^t$. For the derivations in the treatment arm we use $\gamma = (\log(0.8), 1, 0)^t$.

Deviations from base-case

We deviate from the base-case scenario in two ways. First, we alter the overall target settings of sample size, overall treatment effect and prediction model AUC. In a second stage, we consider settings that violate the assumption of a constant relative treatment effect, using a model-based approach.

For the first part, we consider:

- Sample size:
  - $n = 1064$
  - $n = 17000$
- Overall treatment effect:
  - $OR = 0.5$
  - $OR = 1$
- Prediction performance:
  - $AUC = 0.65$
  - $AUC = 0.85$

We set the true risk model coefficients to be $\beta = (-1.63, 0.26, ..., 0.26)^t$ for $AUC = 0.65$ and $\beta = (-2.7, 0.82, ..., 0.82)^t$ for $AUC = 0.85$. In both cases, β_0 is selected so that an event rate of 20% is maintained in the control arm.

For the second part linear, quadratic and non-monotonic deviations from the assumption of constant relative effect are considered. We

also consider different intensity levels of these deviations. Finally, constant absolute treatment-related harms are introduced, i.e. positive ($0.25 \times$ true average benefit), strong positive ($0.50 \times$ true average benefit) and negative ($-0.25 \times$ true average benefit; i.e. constant absolute treatment-related benefit). In case of true absent treatment effects, treatment-related harms are set to $1\%, 2\%$ and -1% for positive, strong positive and negative setting, respectively. The settings for these deviations are defined in Table ***ONLINE TABLE***.

Approaches to individualize benefit predictions

Risk modeling

Merging treatment arms, we develop prediction models including a constant relative treatment effect:

$$P(Y = 1 \mid X = x, Z = z) = g(x^t\beta + \delta_0 z) \tag{B.3}$$

We derive baseline risk predictions for patients by setting $Z = 0$ in @ref(eq:risk). All methods for individualizing benefit predictions are 2-stage methods, that start by fitting a model for predicting baseline risk. The estimated linear predictor of this model is

$$\hat{lp} = lp(x; \hat{\beta}) = x^t\hat{\beta}$$

Risk stratification

Derive a prediction model using the same approach as above and divide the population in equally sized risk-based subgroups. Estimate subgroup-specific absolute benefit from the observed absolute differences. Subject-specific benefit predictions are made by attributing to individuals their corresponding subgroup-specific estimate.

Constant treatment effect

Assuming a constant relative treatment effect, fit the adjusted model in Equation B.3. Then, predict absolute benefit using

$$\hat{\tau}(x; \hat{\beta}, \hat{\gamma}) = g(f(\hat{lp}, 0)) - g(f(\hat{lp}, 1)), \tag{B.4}$$

where $f(\hat{lp}, z) = \hat{lp} + \hat{\delta}_0 z$, with $\hat{\delta}_0$ the estimated relative treatment effect (log odds ratio).

Linear interaction

We relax the assumption of a constant relative treatment effect in Equation B.4 by setting

$$f(\hat{lp}, z) = \delta_0 + \delta_1 z + \delta_2 \hat{lp} + \delta_3 z \hat{lp}$$

Restricted cubic splines

Finally, we drop the linearity assumption and predict absolute benefit using smoothing with restricted cubic splines with $k = 3, 4$ and 5 knots. More specifically, we set:

$$f(\hat{lp}, z) = \delta_0 + \delta_1 z + z s(\hat{lp})$$

where

$$s(x) = \alpha_0 + \alpha_1 h_1(x) + \alpha_2 h_2(x) + \cdots + \alpha_{k-1} h_{k-1}(x)$$

with $h_1(x) = x$ and for $j = 2, \ldots, k-2$

$$h_{j+1}(x) = (x - t_j)^3 - (x - t_{k-1})_+^3 \frac{t_k - t_j}{t_k - t_{k-1}} + (x - t_k)_+^3 \frac{t_{k-1} - t_j}{t_k - t_{k-1}}$$

where $t_1, \ldots, t_k$ are the selected knots[194].

Adaptive model selection frequencies

Discrimination and calibration for benefit

The c-for-benefit represents the probability that from two randomly chosen matched patient pairs with unequal observed benefit, the pair with greater observed benefit also has a higher predicted benefit. To be able to calculate observed benefit, patients in each treatment arm are ranked based on their predicted benefit and then matched 1:1 across treatment arms. Observed treatment benefit is defined as the difference of observed outcomes between the untreated and the treated patient of each matched patient pair. Predicted benefit is defined as the average of predicted benefit within each matched patient pair.

We evaluated calibration in a similar manner, using the integrated calibration index (ICI) for benefit[113]. The observed benefits are regressed on the predicted benefits using a locally weighted scatterplot smoother (loess). The ICI-for-benefit is the average absolute difference between predicted and smooth observed benefit. Values closer to represent better calibration.

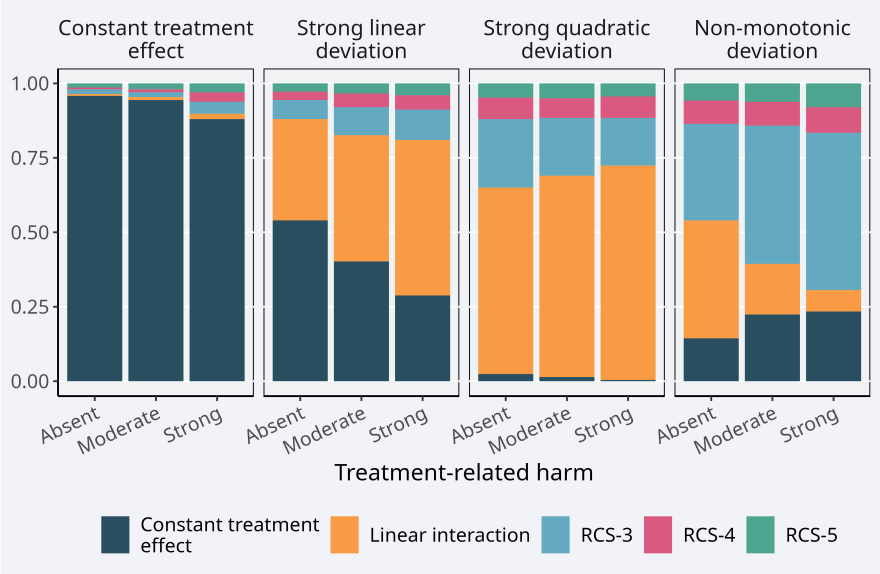

Figure B.1: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. The scenario with the true constant relative treatment effect (first panel) had a true prediction AUC of 0.75 and sample size of 4,250.

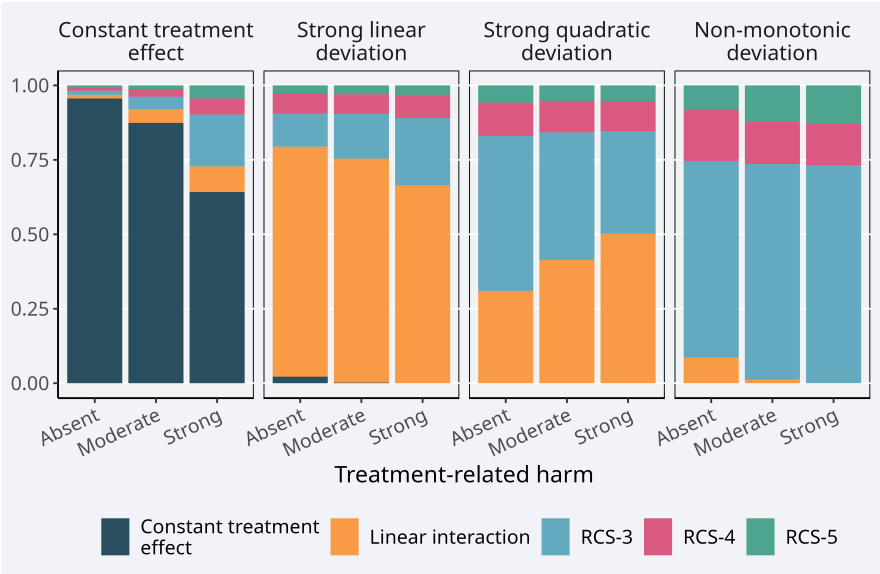

Figure B.2: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. Sample size is 17,000.

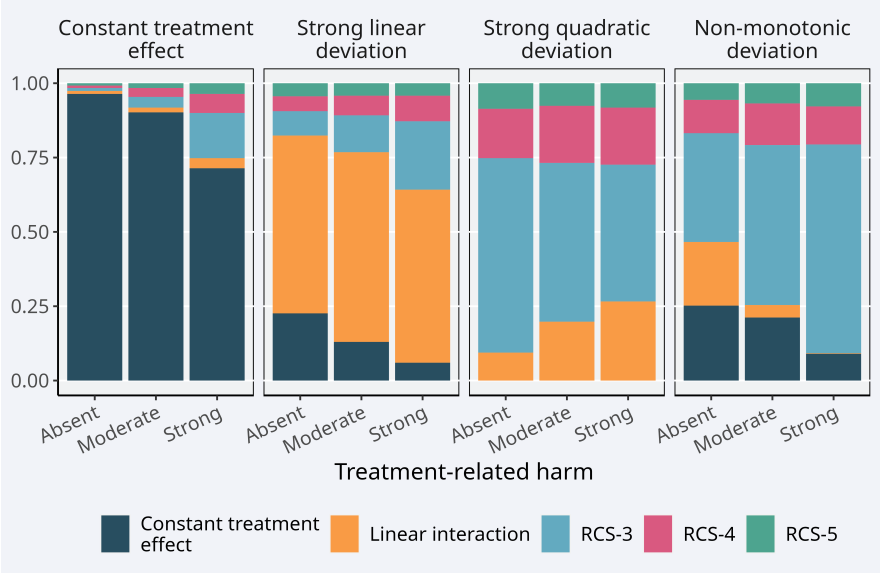

Figure B.3: Model selection frequencies of the adaptive approach based on Akaike's Information Criterion across 500 replications. AUC is 0.85.

Strong relative treatment effect

Here we present the root mean squared error of the considered methods using strong constant relative treatment effect (OR = 0.5) as the reference. Again, the same sample size and prediction performance settings were considered along with the same settings for linear, quadratic and non-monotonic deviations from the base case scenario of constant relative treatment effects are considered. All results can be found at https://arekkas.shinyapps.io/simulation_viewer/.

Treatment interactions

We carried out a smaller set of simulations, in which we assumed true treatment-covariate interactions. Sample size was set to 4,250 and the AUC of the true prediction model was set to 0.75. The following scenarios were considered: 1) 4 true weak positive interactions ($OR_{Z=1}/OR_{Z=0} = 0.83$); 2) 4 strong positive interactions ($OR_{Z=1}/OR_{Z=0} = 0.61$); 3) 2 weak and 2 strong positive interactions; 4) 4 weak negative interactions ($OR_{Z=1}/OR_{Z=0} = 1.17$); 5) 4 strong negative interactions ($OR_{Z=1}/OR_{Z=0} = 1.39$); 6) 2 weak and 2 strong negative interactions; 7) combined positive and negative strong interactions. We also considered constant treatment-related harms applied on

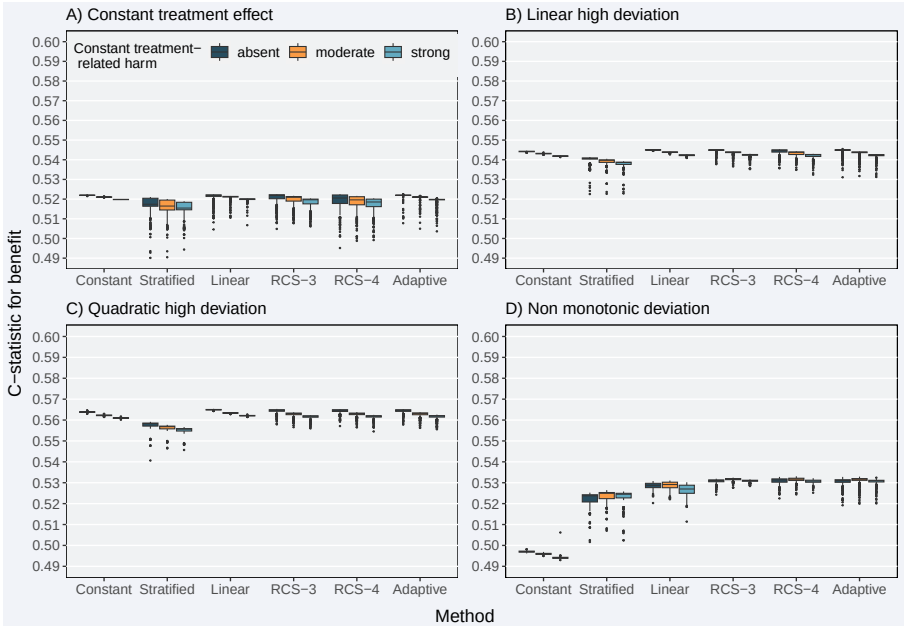

Figure B.4: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 17,000.

the absolute scale to all treated patients. The exact settings were: 1) absent treatment-related harms; 2) moderate treatment-related harms, defined as 25% of the average true benefit of the scenario without treatment-related harms; 3) strong treatment-related harms defined as 50% of the true average benefit of the scenario without treatment-related harms; 4) negative treatment-related harms (benefit), defined as an absolute risk reduction for treated patients of 50% of the true average benefit of the scenario without treatment-related harms. The exact settings can be found in Table ***ONLINE TABLE***.

Correlated covariates

We analyzed the sensitivity of our results to correlation between baseline characteristics by incuding additional simulation scenarios. We sampled covariates $W_1, \dots, W_8 \sim N(0, \Sigma)$. We generated four continuous baseline covariates $X_1 = W_1, \dots, X_4 = W_4$ and four binary covariates with 20% prevalence $X_5 = I(W_5 > z_{0.8}), X_8 = I(W_8 > z_{0.8})$, where I is the indicator function. We selected the covariance matrix Σ such that $\text{cor}(X_i, X_j) = 0.5$, for any $i \neq j$. More precisely, we set Σ as can be seen below:

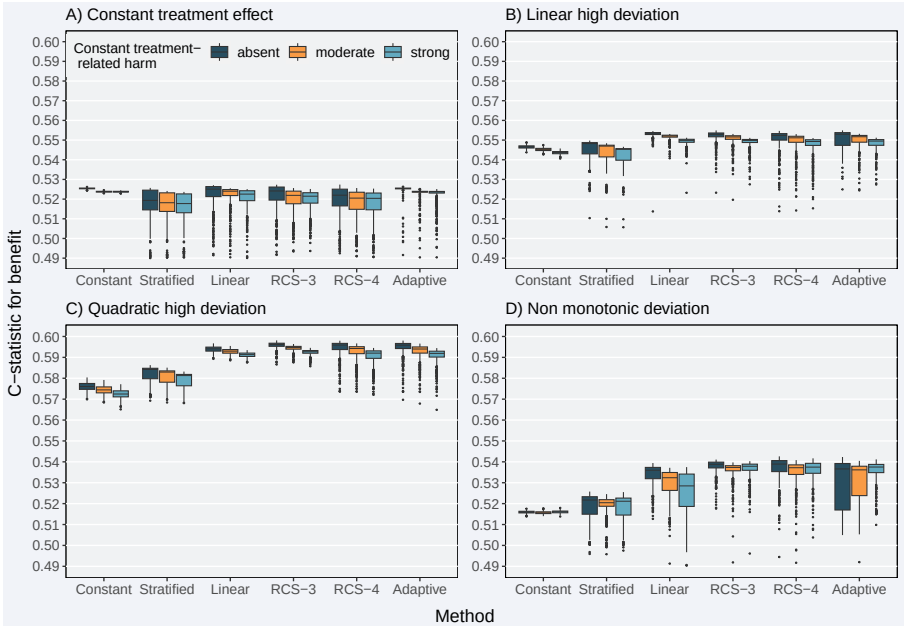

Figure B.5: Discrimination for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.85 and sample size of 4,250.

$$\Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.708 & 0.708 & 0.708 & 0.708 \\ 0.5 & 1 & 0.5 & 0.5 & 0.708 & 0.708 & 0.708 & 0.708 \\ 0.5 & 0.5 & 1 & 0.5 & 0.708 & 0.708 & 0.708 & 0.708 \\ 0.5 & 0.5 & 0.5 & 1 & 0.708 & 0.708 & 0.708 & 0.708 \\ 0.708 & 0.708 & 0.708 & 0.708 & 1 & 0.745 & 0.745 & 0.745 \\ 0.708 & 0.708 & 0.708 & 0.708 & 0.745 & 1 & 0.745 & 0.745 \\ 0.708 & 0.708 & 0.708 & 0.708 & 0.745 & 0.745 & 1 & 0.745 \\ 0.708 & 0.708 & 0.708 & 0.708 & 0.745 & 0.745 & 0.745 & 1 \end{pmatrix}$$

In order to ensure that the simulated datasets were comparable to the original main simulation scenarios, i.e. control arm event rate of 20% and true risk model c-statistic (AUC) equal to the target, we needed to adjust the coefficients of the true risk model. The exact settings of the simlation scenarios for the sensitivity analyses can be found in Table ***ONLINE TABLE***.

We found no noticeable differences between methods for individualizing treatment benefit predictions compared to the results of the simulation scenarios where baseline covariates were assumed to be statistically independent (Fig-

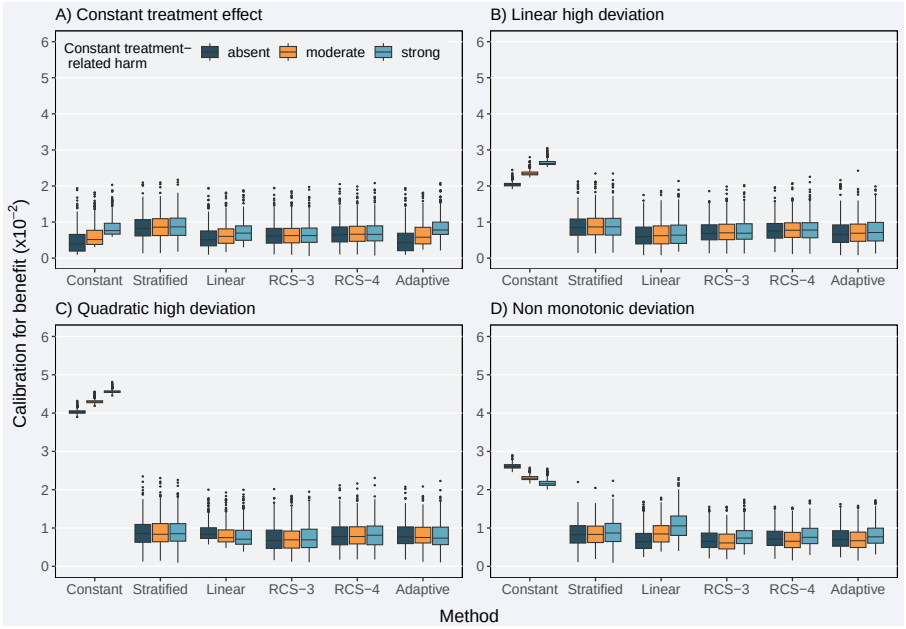

Figure B.6: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.75 and sample size of 17,000.

ures B.11 to B.13).

Empirical illustration

For predicting baseline risk of 30-day mortality we fitted a logistic regression model with age, Killip class (*Killip*), systolic blood pressure (*sysbp*), pulse rate (*pulse*), prior myocardial infarction (*pmi*), location of myocardial infarction (*miloc*) and treatment as the covariates. Baseline predictions were made setting treatment to 0.

$$P(outcome = 1|X = x) = \text{expit}(lp(x)), (\#eq : gusto1)$$

where

$$\begin{aligned} lp(x) =& \beta_0 + \beta_1 \text{age} + \beta_2 I(\text{Killip} = II) + \beta_3 I(\text{Killip} = III) + \\ & \beta_4 I(\text{Killip} = IV) + \beta_5 min(\text{sysbp}, 120) + \beta_6 \text{pulse} + \\ & \beta_7 max(\text{pulse - 50, 0}) + \beta_8 I(\text{pmi} = yes) + \\ & \beta_9 I(\text{miloc} = Anterior) + \beta_9 I(\text{miloc} = Other) + \\ & \gamma \times \text{treatment} \end{aligned} \qquad (\#eq : gusto2)$$

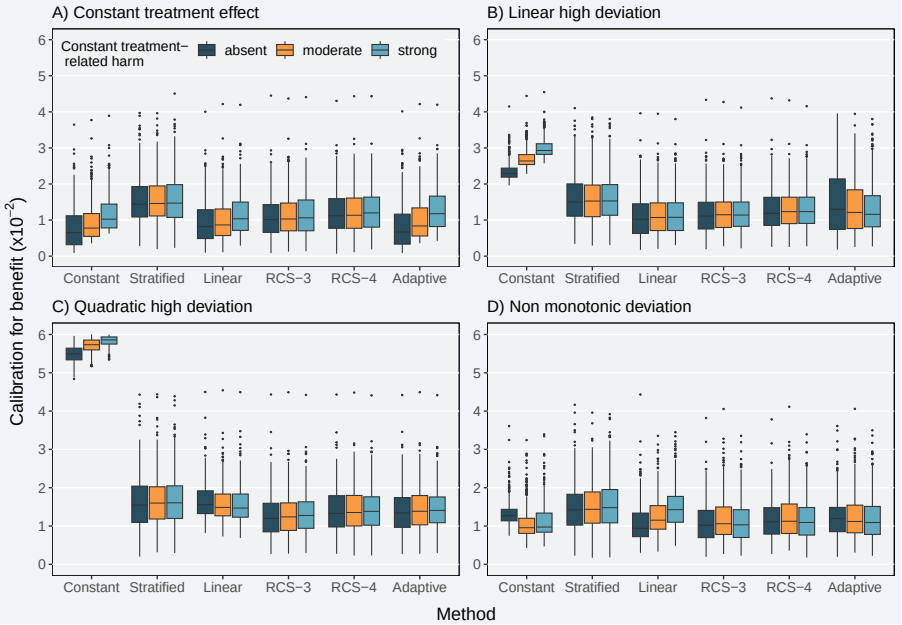

Figure B.7: Calibration for benefit of the considered methods across 500 replications calculated in a simulated sample of size 500,000. True prediction AUC of 0.85 and sample size of 4,250.

and $expit(x) = \frac{e^x}{1+e^x}$

Table B.1: Coefficients of the prediction model for 30-day mortality, based on the data from GUSTO-I trial.

Variable	Estimate	stderror	zvalue	pvalue
Intercept	-3.020	0.797	-3.788	0.000
Age	-0.208	0.053	-3.935	0.000
Killip class = II	0.077	0.002	31.280	0.000
Killip class = III	0.614	0.059	10.423	0.000
Killip class = IV	1.161	0.121	9.566	0.000
Systolic blood pressure	1.921	0.162	11.872	0.000
Pulse rate (1)	-0.039	0.002	-20.332	0.000
Pulse rate (2)	-0.024	0.016	-1.521	0.128
Previous MI (yes)	0.043	0.016	2.675	0.007
MI location (Other)	0.447	0.056	7.964	0.000
MI location (Anterior)	0.286	0.135	2.126	0.033
Treatment	0.543	0.051	10.625	0.000

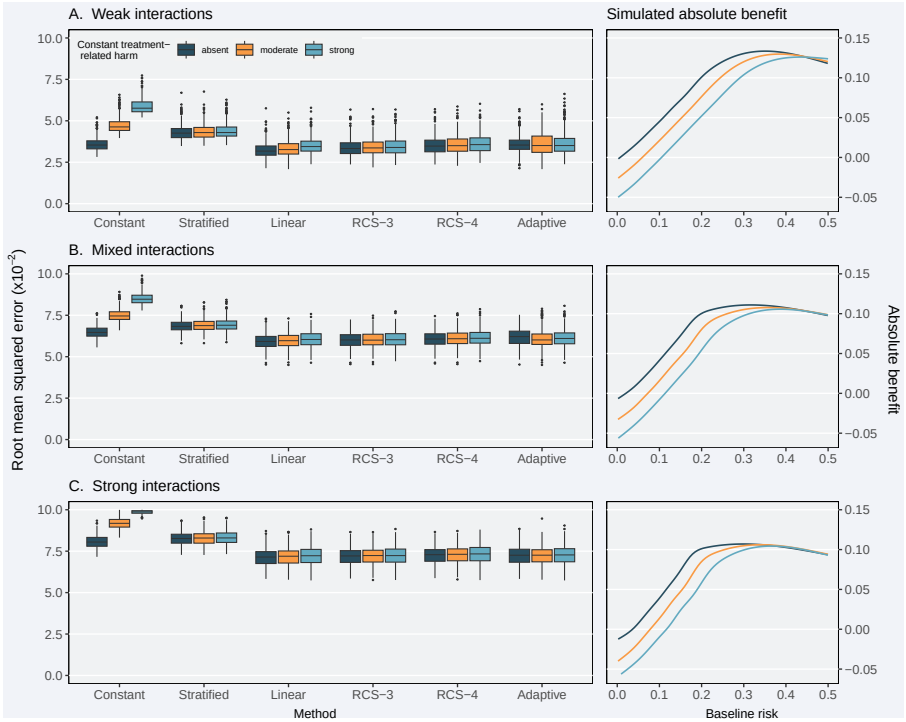

Figure B.8: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions all favoring treatment were considered.

Bootstrap confidence intervals

Bootstrap confidence intervals in Figure 6 of the main manuscript were derived using the following approach:

1. Draw with replacement a sample D^* from the original dataset D of the same size as D.
2. Fit a logistic regression model m^* in D^* to predict 30-day mortality using the same covariates as the initial model m estimated in D.
3. Using the linear predictor $\widehat{lp}^*$ of model m^*, fit models for constant relative treatment effect, linear interaction of the linear predictor with treatment, and interaction of treatment with a restricted cubic splines transformation of the linear predictor in D^*, as described in the *Methods* section.
4. Based on m^*, use predicted risks to stratify D^* into risk quarters and estimate absolute treatment effects within risk strata.

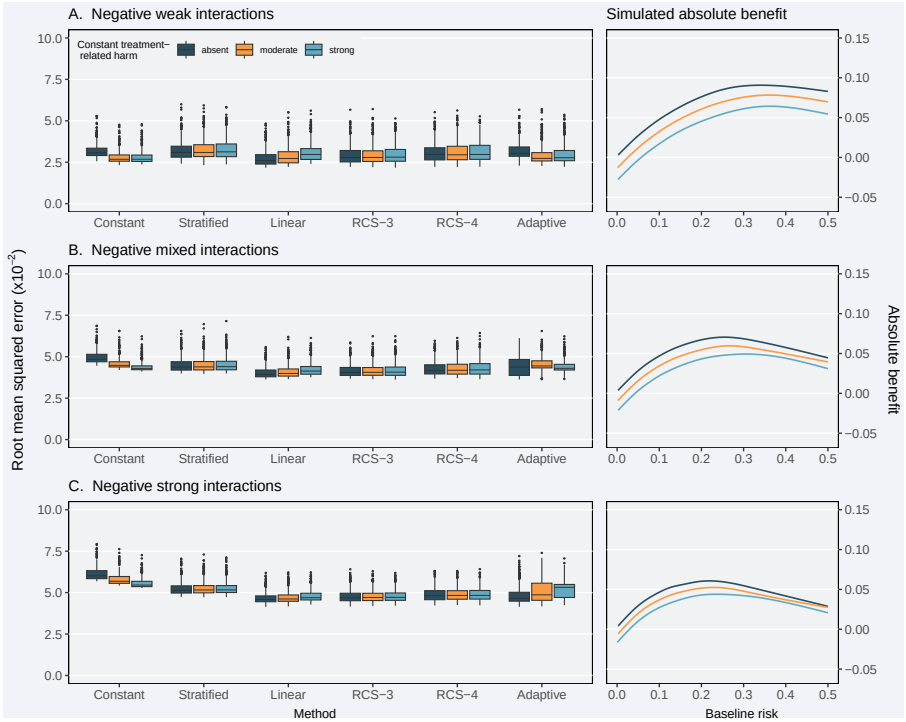

Figure B.9: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions all favoring the control were considered.

5. Repeat steps 1 through 4 for a total of $b = 10000$ times.
6. Derive the confidence band for each continuous method (not risk stratification) as follows:
   a. Split the range $(0, 0.25]$ of baseline risk values into 499 equal-length intervals of the form $(p_0, p_1], \ldots, (p_{498}, p_{499}]$.
   b. For a specific point $p_k, k = 1 \ldots, 499$ and each method i, use the 2.5 and 97.5 percentiles of the b absolute benefit estimates to define the confidence interval $(q^{i,p_k}_{0.025}, q^{i,p_k}_{0.975})$.
7. Derive two sets of confidence intervals for the risk stratification approach as follows:
   a. *Treatment effect:* For each risk quarter identified by the original model m developed on D, the confidence interval for the mean absolute treatment effect is derived from the 2.5 and 97.5 percentiles of the mean absolute treatment effects estimated within each risk

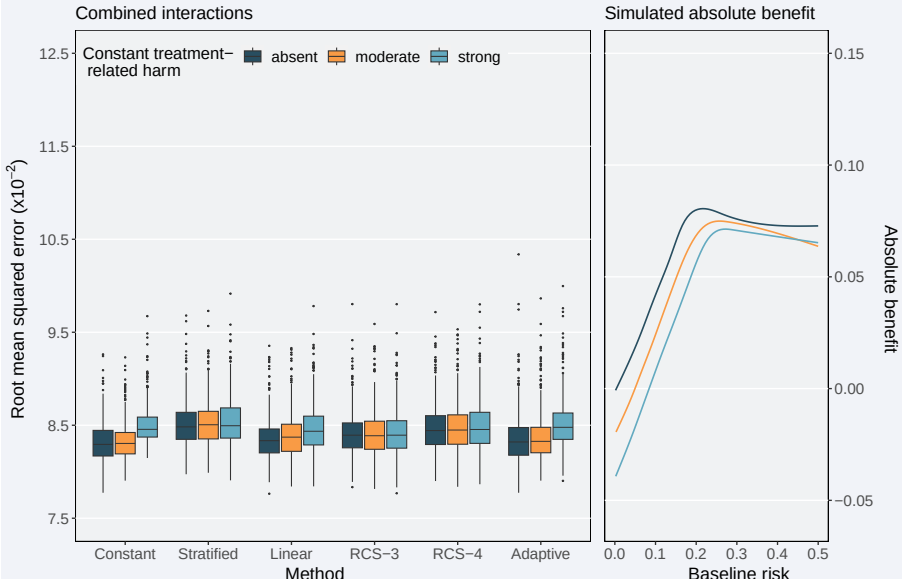

Figure B.10: RMSE of the considered methods across 500 replications calculated in a simulated sample of size 500,000 where treatment-covariate interactions 2 favoring treatment and 2 favoring the control were considered.

quarter across the b bootstrap samples.

b. *Mean predicted risk:* For each risk quarter identified by the original model m, the confidence interval for the quarter-specific mean predicted risk is derived from the 2.5 and 97.5 percentiles of the mean predicted risks estimated within each risk quarter across the b bootstrap samples.

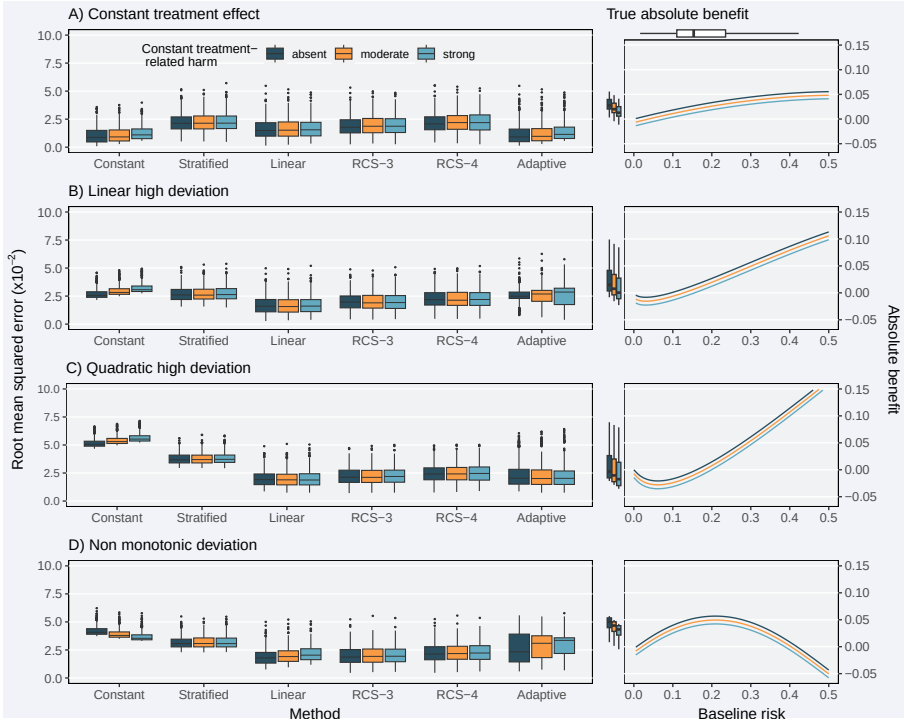

Figure B.11: RMSE of the sensitivity analyses assuming correlated baseline covariates. All considered methods were derived using simulated samples of size 4,250 and true prediction c-statistic of 0.75.

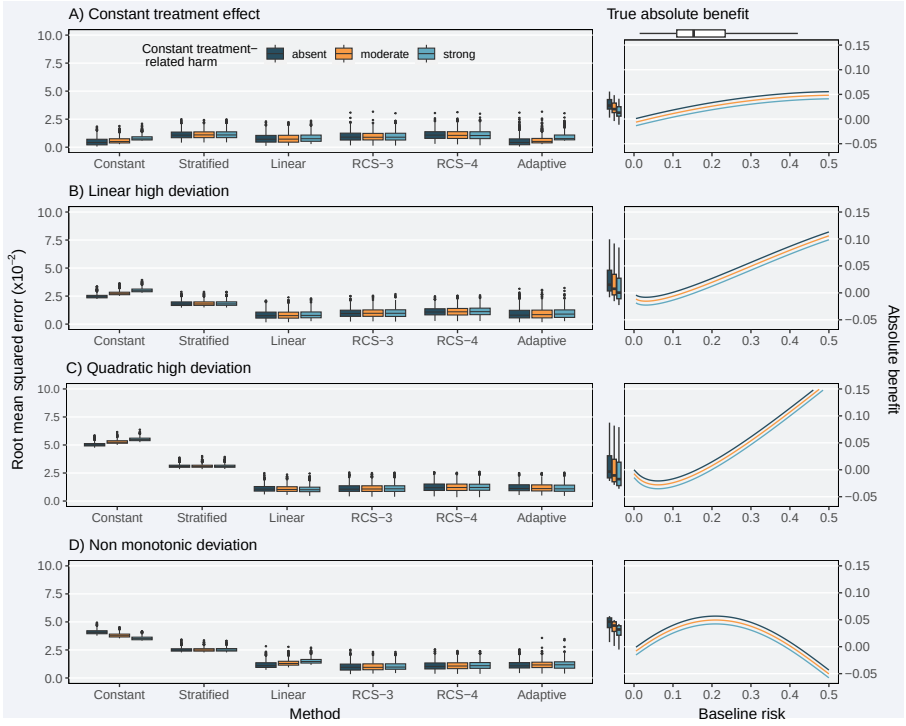

Figure B.12: RMSE of the sensitivity analyses assuming correlated baseline covariates. All considered methods were derived using simulated samples of size 17,000 and true prediction c-statistic of 0.75.

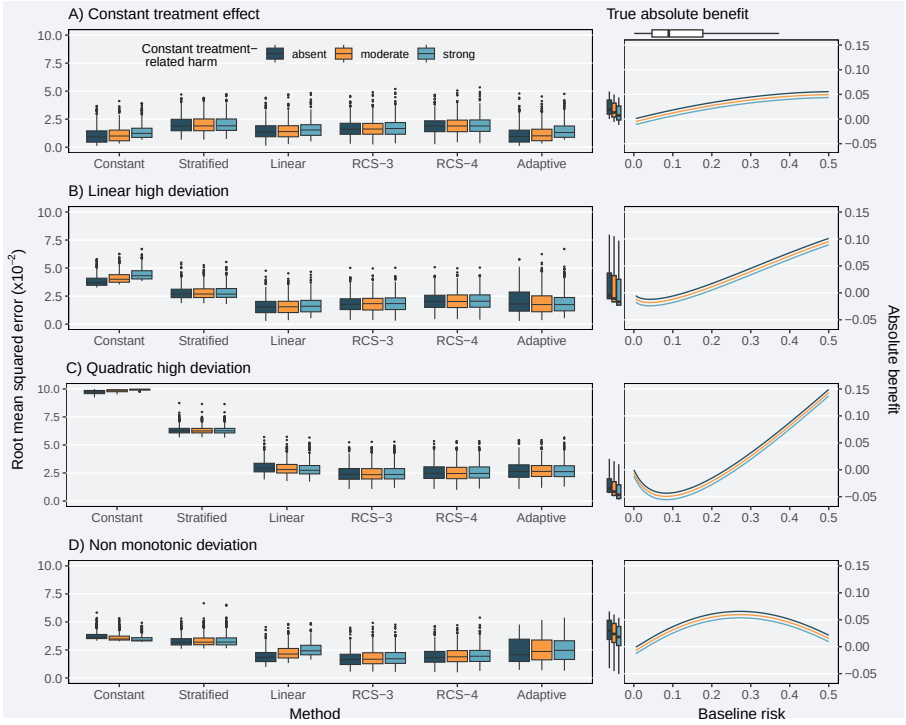

Figure B.13: RMSE of the sensitivity analyses assuming correlated baseline covariates. All considered methods were derived using simulated samples of size 4,250 and true prediction c-statistic of 0.85

Summary

In this thesis we focus on baseline risk and its use for guiding medical decision making. Development of risk prediction models, i.e. mathematical functions relating the presence of the outcome of interest to a set of measured predictors (covariates), is crucial for personalizing outcome risk and, consequently, individualizing treatment decisions.

Overall results from randomized controlled trials or large observational studies often do not apply to individual patients and represent a summary of heterogeneous individual treatment effects. If variation of these individual effects is large, then reliance on overall estimates may result in sub-optimal decisions. Evaluation of treatment effect heterogeneity has been a focal point of methods research in recent years. To this end, we systematically reviewed the literature on predictive approaches for the evaluation of treatment effect heterogeneity.

Baseline risk is an important determinant of treatment effect and, therefore, can be directly used to predict individualized treatment benefits. Consequently, we developed and compared methods for risk-based assessment of treatment effect heterogeneity both on the randomized controlled trial and the observational setting. Finally, in applications we used risk-based methods to better guide medical decisions in the fields of melanoma, COVID-19, and osteoporosis treatment.

In Chapter 1 we presented the results of a literature review using a broad search strategy, complemented by suggestions from a technical expert panel. We classified the identified approaches into three categories (risk-based, treatment effect modeling, and optimal treatment regime methods). Risk-based methods use only prognostic factors to define patient subgroups, relying on the mathematical dependency of the absolute risk difference on baseline risk. Treatment effect modeling methods use both prognostic factors and treatment effect modifiers to explore characteristics that interact with the effects of therapy on a relative scale. Finally, optimal treatment regime methods focus primarily on treatment effect modifiers to classify the trial population into those who benefit from treatment and those who do not.

In Chapter 2 we presented a standardized framework for the evaluation of treatment effect heterogeneity using a risk-based approach. The proposed framework consists of five steps: 1) definition of the research aim, i.e., the population, the treatment, the comparator and the outcome(s) of interest; 2) identification of relevant databases; 3) development of a prediction model for the outcome(s) of interest; 4) estimation of relative and absolute treatment effect within strata of predicted risk, after adjusting for observed confounding; 5) presentation of the results. We demonstrated our framework by evaluating heterogeneity of the effect of thiazide or thiazide-like diuretics versus angiotensin-converting enzyme inhibitors on three efficacy and nine safety outcomes across three observational databases. We showed that patients at low risk of acute myocardial infarction receive negligible absolute benefits for all three efficacy outcomes, though these were more pronounced in the highest risk group, especially for acute myocardial infarction.

In Chapter 3 we presented the results of an extensive simulation study for the comparison of easily applicable risk-based methods for the prediction of individualized treatment effects in the setting of randomized controlled trials. More specifically, we compared models with a constant relative treatment effect, models including a linear interaction of treatment with the prognostic index of baseline risk, and models including an interaction of treatment with restricted cubic spline transformation of the prognostic index. We also considered an adaptive approach using Akaike's information criterion for automatically selecting among the previous methods. We showed that the linear-interaction model has optimal or close-to-optimal performance across many simulation scenarios with moderate sample size. The restricted cubic splines model required strong non-linear deviations from a constant treatment effect larger sample size. We also applied these methods in actual data using the dataset of the GUSTO-I trial.

In Chapter 4 we developed prediction models for the management of patients with sentinel node-positive melanoma. We first developed a model for recurrence, which we then re-calibrated for the prediction of distant metastasis and overall mortality, allowing for the prediction of all three outcomes from the same model with adequate accuracy (AUC of 0.68, 0.70, and 0.70 for recurrence, distant metastasis, and overall mortality, respectively). An important addition of these models is that they do not require information on positive lymph nodes after completion lymph node dissection, which is no longer routine practice for sentinel node-positive melanoma patients. Finally, we provided a nomogram for graphical presentation of our derived prediction models.

In Chapter 5 we developed simple and valid models for predicting mortality

and need for intensive care unit admission in patients presenting at the emergency department with suspected COVID-19. We used first-wave patients from March till August 2020 for model development and second-wave patients from September till December 2020 for model validation. The final model for predicting mortality was based on age and logarithmic transforms of respiratory rate, C-reactive protein, lactate dehydrogenase, albumin, and urea. For the prediction of admission to the intensive care unit we re-calibrated the mortality prediction model. Our overall mortality prediction model displayed good discrimination and calibration across all hospitals in the development dataset (AUC in 4 hospitals 0.85 with 95% CI 0.81 to 0.88; 0.81 with 95% CI 0.71 to 0.91; 0.86 with 95% CI 0.82 to 0.90; 0.85 with 95% CI 0.81 to 0.88), while it also maintained good performance at temporal validation analysis (AUC in four hospitals: 0.82 with 95% CI 0.78 to 0.86; 0.82 with 95% CI 0.74 to 0.90; 0.79 with 95% CI 0.70 to 0.88; 0.83 with 95% CI 0.79 to 0.86). The resulting COPE models were implemented as a publicly accessible web-based application and as independent mobile applications which also included detailed description of the derivation of COPE, descriptions of the derivation data and reported on model performance to ensure transparency.

In Chapter 6…

In conclusion, we separated methodological approaches to the assessment of heterogeneity of treatment effect in randomized controlled trials into three categories, these are, risk-based approaches, treatment effect modeling approaches, and optimal treatment regime approaches. We successfully developed and implemented a risk-based framework for the assessment of heterogeneity of treatment effect in the observational setting. We showed, through extensive simulations, that regression models with a linear interaction of baseline risk with treatment provide a viable option for the prediction of personalized treatment benefits with smaller sample sizes or moderately performing risk prediction models. We showed that prediction models developed in chapters Chapter 4 and Chapter 5 can be used to aid medical decisions in a timely manner. Finally, the implementation of our standardized framework for risk-based assessment of heterogeneity of treatment effect in the field of osteoporosis uncovered the potential of our methodology, while it also demonstrated its limitations due to the observational nature of the data.