A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases

Alexandros Rekkas, MSc[1] David van Klaveren, PhD[2,3], Patrick B. Ryan, PhD[4]
Ewout W. Steyerberg, PhD[3,5] David M. Kent, PhD[2] Peter R. Rijnbeek, PhD[1]

[1] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands

[2] Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center, Boston, MA, USA

[3] Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands

[4] Janssen Research and Development, 125 Trenton Harbourton Rd, Titusville,NJ 08560, USA

[5] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Corresponding author

Alexandros Rekkas, MSc
Department of Medical Informatics
Erasmus University Medical Center
3000 CA Rotterdam, P.O. Box 2040
Email: a.rekkas@erasmusmc.nl

Funding

This work has been performed in the European Health Data and Evidence Network (EHDEN) project. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Abstract

One of the aims of the Observational Health Data Sciences and Informatics (OHDSI) initiative is population-level treatment effect estimation in large observational databases. Since treatment effects are well-known to vary across groups of patients with different baseline risk, we aimed to extend the OHDSI library of open-source tools with a framework for risk-based assessment of treatment effect heterogeneity. The proposed framework consists of five steps: 1) definition of the problem, 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 propensity scores within strata of predicted risk and estimation of relative and absolute treatment effect within strata of predicted risk; 5) evaluation and presentation of results. We demonstrate our framework by evaluating heterogeneity of the effect of angiotensin-converting enzyme (ACE) inhibitors versus beta blockers on a set of 9 outcomes of interest across three observational databases. With increasing risk of acute myocardial infarction we observed increasing absolute benefits, i.e. from -0.03% to 0.54% in the lowest to highest risk groups. Cough-related absolute harms decreased from 4.1% to 2.6%. The proposed framework may be useful for the evaluation of heterogeneity of treatment effect on observational data that are mapped to the OMOP Common Data Model. The proof of concept study demonstrates its feasibility in large observational data. Further insights may arise by application to safety and effectiveness questions across the global data network.

Keywords: observational data, heterogeneity of treatment effect, risk stratification, subgroup analysis

1 Introduction

Understanding how a treatment's effect varies across patients—a concept described as heterogeneity of treatment effects (HTE)—has been central to the agenda for both personalized (or precision) medicine and comparative effectiveness research. More formally, HTE has been defined as the non-random variability in the direction or magnitude of a treatment effect, in which the effect is measured using clinical outcomes [1]. Usually, analyses focus on the relative scale, where treatment effects are assessed one at a time in patient subgroups defined from single covariates, an approach that suffers from low power and multiplicity issues [2,3]. However, even with well-established constant relative effects, treatment benefit (or harm) may vary substantially on the absolute scale.

In recent years, a large number of methods has been developed for the assessment of HTE, mainly in the RCT setting. Earlier work suggested separating HTE analyses into exploratory, confirmatory, descriptive and predictive [4]. Exploratory analyses focus on hypothesis generation, confirmatory analyses test subgroup effect hypotheses, descriptive analyses aim at facilitating future synthesis of subgroup effects and predictive analyses predict probabilities of benefit or harm in individual patients. Predictive HTE approaches can be further subdivided into risk modeling, treatment effect modeling and optimal treatment regime methods, based on the reference class used for defining patient similarity when making individualized predictions or recommendations [5]. We focus on "risk modeling" approaches where patients are divided into risk strata using either an existing or an internally developed risk prediction model. Risk-stratum-specific estimates provide an overview of the evolution of treatment effects with increasing risk both on the relative and the absolute scale. Recently, systematic guidance on the application of such methods has been developed [6,7].

While these approaches were developed for application in randomized controlled trials (RCTs), observational databases are also an appealing substrate. Observational healthcare databases, such as administrative claims and electronic health records, are already highly available for the analysis of pharmacoepidemiologic research questions [8,9]. They are also often larger than many typical trials, providing excellent power for HTE analysis, including heterogeneous populations. However, unlike RCTs, treatment effects are subject to confounding, while the unique structure of different databases calls for database-specific analysis plans that often are not easily transportable. Because of the latter issue, running analyses at scale demands a big investment of time and effort, as researchers are forced to map their analysis plans to the databases available to them.

The Observational Health Data Sciences and Informatics (OHDSI) collaborative has established an international network of data partners and researchers that aim to bring out the value of health data through large-scale analytics by mapping all available databases to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [10,11].

We aimed to develop a framework for implementing a risk-based predictive approach for evaluating HTE in high-dimensional observational data, extending the existing guidelines of the RCT setting. Our publicly available package provides an out-of-the-box solution for implementing such analyses at scale within the OHDSI network, taking advantage of the OMOP-CDM. We implemented the framework using existing OHDSI methods including the patient-level prediction framework and the population-level effect estimation framework based on new-user cohort design [12,13]. As a proof-of-concept we analyzed heterogeneity of the effects of first-line hypertension treatment: we compared the effect of angiotensin converting enzyme (ACE) inhibitors to beta blockers on 9 outcomes across three different US claims databases.

2 Materials and Methods

The proposed framework defines 5 distinct steps that enable a standardized approach for risk-based assessment of treatment effect heterogeneity for databases mapped to the OMOP-CDM. These are: 1) general definition of the research aim; 2) identification of the database within which the analyses will be performed; 3) a prediction step where internal or external prediction models are used to assign patient-level risk predictions; 4) an estimation step where absolute and relative treatment effects are estimated 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. The source code can be found at https://github.com/OHDSI/RiskStratifiedEstimation.

2.1 Step 1: General definition of the problem

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

Cohort definitions are crucial for this step of the framework. We define a cohort as the set of patients who satisfy one or more inclusion criteria for a duration of time. A cohort within the OHDSI setting is more than a set of specific clinical codes, providing a definition of a logic for how to use that code set. All cohort definitions consist of: an entry event, i.e. the time a patient enters a cohort; a set of inclusion criteria applied to the initial event cohort to further restrict the set of people, resulting in the construction of the construction of the qualifying cohort; cohort exit criteria that terminate the patient's presence in the cohort. Cohort definitions are transportable, meaning that in theory they can be implemented in any database, provided that it is mapped to the OMOP-CDM.

Our framework uses a comparative cohort design. This means that at least 3 cohorts of patients need to be defined at this stage of the framework.

- A single treatment cohort (T) which includes patients with disease D receiving the target treatment of interest. For example, a set of hypertension patients within a database that receive angiotensin-converting enzyme inhibitors, followed from the time of initiation until the time of censoring.
- A single comparator cohort (C) which includes patients with disease D receiving the comparator (control) treatment. For example, a set of patients in a database that receive beta blockers, followed from the time of initiation until the time of censoring.
- One or more outcome cohorts ($O_1, \ldots, O_n$) that contain patients developing the outcomes of interest. For example, the set of patients in a database that have at least one occurrence of acute myocardial infarction (MI) in their record.

2.2 Step 2: Identification of the database

The aim of this step is the inclusion of databases that represent the patient population of interest. It is required that the databases are mapped to the OMOP-CDM. The inclusion of multiple databases 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 issues that may be of importance for database inclusion are the depth of data capture (the precision at which measurements, lab tests, conditions are recorded), the reliability of data entry and many more, also depending on the task at hand.

2.3 Step 3: Prediction

We adopt the standardized framework for the generation of patient-level prediction models using observational data that ensures adherence to existing guidelines [14,15]. This prediction framework requires the definition of two essential cohorts: a target cohort, i.e. a set of patients that satisfy one or more inclusion criteria for a duration of time, and an outcome cohort.

To generate the target cohort we pool the already defined treatment cohort and comparator cohort. Further restrictions can be applied on the target cohort to construct the final population on which the prediction model will be developed (e.g. exclude patients with a prior outcome in their history, before being included in the target cohort). To avoid deferentially fitting the prediction model to patients across treatment arms, thus introducing spurious interactions with treatment [16,17], we develop the patient-level prediction model in the propensity score-matched (1:1) subset of the population.

More specifically, we first estimate propensity scores using LASSO logistic regression and a large set of baseline covariates including demographics, drug exposures, diagnoses, measurements and medical devices. We match

patients 1-1 using a caliper, i.e. the maximum distance that is acceptable for any match. The default value we use is 0.2 on the standardized logit scale for the propensity scores. Other methods of fitting the propensity scores, such as random forest and others can also be considered.

Finally, we need to define the time horizon within which we aim to make predictions and we also need to select the machine-learning algorithm we want to use to generate patient-level predictions. Currently, the available options are regularized logistic regression, random forest, gradient boosting machines, decision tree, naive Bayes, K-nearest neighbors, neural network and deep learning (convolutional neural networks, recurrent neural network and deep nets).

After model development, a performance overview of the derived prediction models including discrimination and calibration both in the propensity score matched subset, the entire population and separately for treated and comparator patients should also be reported. This is important to ensure that no overfitting of the prediction model in one of the cohorts has occurred. In addition, the performance of the prediction models is directly related to our ability to single out patient subgroups where treatment may be highly beneficial or unsafe. Kent et al [18] demonstrated that the event rate and the discriminative ability of the prediction model can predict very well the distribution of predicted risk. Lower event rate and higher c-statistic (given good calibration) result in high risk heterogeneity, thus making estimated average treatment effects uninformative. In this case, risk stratified analysis of HTE can be more effective in singling out patient subgroups that stand to benefit (or be harmed) most by treatment in question.

2.4 Step 4: Estimation

The aim of this step is the estimation of treatment effects (both on the relative and the absolute scale) within risk strata—typically 4 risk quarters—defined using the prediction model of step 3. 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 evaluation of relative and absolute treatment effects can be considered, as long as the this is done consistently in all risk strata. Common approaches 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 stratify them into groups with similar propensity scores or to weigh each patient's contribution to the estimation process [19].

Before focusing on the results of the estimation process we need to evaluate if adequate covariate balance was achieved within each risk stratum accounting for measured confounding. Common approaches include evaluation of the overlap of propensity score distributions and calculation of standardized covariate differences before and after propensity score adjustment.

A schematic overview of the prediction and estimation steps is presented in Figure 1.

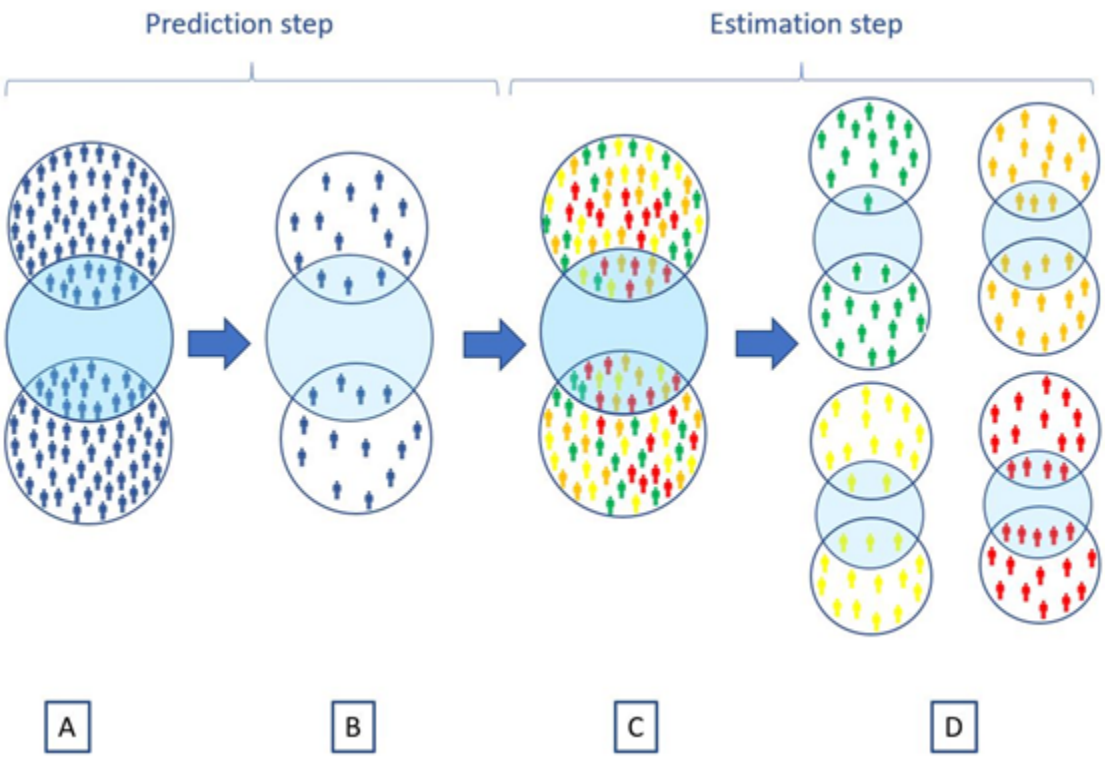

Figure 1: (A) Starting from a treatment (top), a comparator (bottom) and an outcome (middle) cohort we estimate the propensity scores on the entire target population. (B) We match patients on the propensity scores and estimate the prediction model. Since we match patients we develop the prediction model on smaller subset of the initial population and, therefore, the number of patients is smaller in B compared to A. (C) We apply the prediction model on the entire population (green: lower 25% of the risk distribution; yellow: patients with risk between 25% and 50% of the risk distribution; orange: patients with risk between 50% and 75% of the risk distribution; red: patients at risk higher than 75% of the risk distribution). (D) We separate in risk subgroups, here quarters. Within risk quarters propensity scores are estimated again and relative and absolute treatment effects are estimated.

2.5 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. We find that a side-by-side presentation on a forest-like format can give a very good representation of our results.

3 Results

As a proof of concept, we focus on the comparison of angiotensin converting enzyme (ACE) inhibitors to beta blockers are among the most common treatment classes for hypertension, with well-established effectiveness. Beta blockers, even though initially widely used for the treatment of hypertension, more recent trials and meta-analyses have cast doubt on their relative effectiveness [20]. As a result, newer US guidelines do not consider beta blockers for initial treatment for hypertension while in the EU guidelines combination with other antihypertensive treatments is recommended [21,22]. However, another meta-analysis suggested that the efficacy profile of beta blockers is similar to other major treatment classes in younger hypertensive patients and, thus, countries like Canada still include them as a first-line treatment candidate [23,24].

3.1 Step 1: General definition of the problem

We consider the following research aim: "compare the effect of ACE-inhibitors (T) to the effect of beta blockers (C) in patients with established hypertension (D) with respect to 9 outcomes ($O_1, \ldots, O_9$)". The cohorts are:

- Treatment 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.
- Comparator cohort: Patients receiving any drug within the beta blocker class with at least one year of follow-up before treatment initiation and a recorded hypertension diagnosis within that year.
- Outcome cohorts: We consider 3 main and 6 safety outcome cohorts. These are patients in the database with a diagnosis of: acute MI; hospitalization with heart failure; ischemic or hemorrhagic stroke (efficacy outcomes); hypokalemia; hyperkalemia; hypotension; angioedema; cough; abnormal weight gain (safety outcomes).

All cohort definitions were identical to the ones used in the multinational study carried out within OHDSI that provided overall treatment effect estimates comparing all anti-hypertensive drug classes with each other [25]. More information can be found in the supplementary material.

3.2 Step 2: Identification of the databases

We used the following databases:

- IBM MarketScan Medicare Supplemental Beneficiaries (MDCR): Represents health services of retirees (aged 65 or older) in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service or capitated health plans. These data include adjudicated health insurance claims (e.g. inpatient, outpatient and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives.
- IBM MarketScan Medicaid (MDCD): Adjudicated US health insurance claims for Medicaid enrollees from multiple states. It includes hospital discharge diagnoses, outpatient diagnoses and procedures and outpatient pharmacy claims as well as ethnicity and Medicare eligibility.
- IBM MarketScan Commercial Claims and Encounters (CCAE): Data from individuals enrolled in US employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses and dependents. Additionally, it captures laboratory tests for a subset of the covered lives.

Our analyses included a total of 784,561, 66,820 and 101,661 patients initiating treatment with ACE inhibitors and 395,740, 45,999 and 69,798 patients initiating treatment with beta blockers in CCAE, MDCD and MDCD respectively (Table 1). Adequate numbers of patients were included in all strata of predicted acute MI risk.

Table 1: Number of patients, person years and events within quarters of predicted risk for acute MI for the 3 main outcomes of the study (acute MI, hospitalization with heart failure and ischemic or hemorrhagic stroke) across the 3 databases. The populations used for the evaluation of hosptialization with heart failure and stroke are subsets of the populations used for the evaluation of actute MI.

		ACE inhibitors			Beta blockers		
Outcome	Risk quarter	Persons	Person years	Events	Persons	Person years	Events
CCAE							
Acute myocardial infarction	1	161,099	276,171	203	133,977	220,633	135
	2	204,882	372,197	534	90,193	169,231	321
	3	214,413	393,583	117	80,662	150,035	535
	4	204,167	351,727	2,095	90,908	154,419	1,520
Heart failure (hosp)	1	146,259	249,809	228	126,387	206,706	378

	2	188,006	341,014	457	84,280	158,425	340
	3	218,052	399,394	826	83,421	155,222	570
	4	230,226	400,330	2,012	98,380	169,139	1,773
Stroke (ischemic or hemorrhagic)	1	146,069	294,484	299	126,264	206,453	320
	2	187,524	340,234	554	84,000	157,913	351
	3	217,070	397,830	947	83,038	154,587	521
	4	226,128	393,861	1,718	97,628	167,810	1,077
MDCD							
Acute myocardial infarction	1	14,347	19,972	15	13,858	17,056	20
	2	18,412	26,737	99	9,793	14,180	53
	3	18,893	31,231	226	9,312	15,041	174
	4	15,168	27,383	561	13,036	22,158	587
Heart failure (hosp)	1	18,004	25,006	87	16,028	20,042	120
	2	18,190	27,253	208	9,108	13,527	138
	3	17,386	29,261	453	8,618	14,219	340
	4	11,775	21,440	970	9,928	17,197	1,155
Stroke (ischemic or hemorrhagic)	1	17,963	24,939	59	15,996	19,991	46
	2	18,063	27,086	180	9,045	13,411	104
	3	17,129	28,846	356	8,582	14,155	208
	4	11,917	21,627	536	10,891	18,601	573
MDCR							
Acute myocardial infarction	1	27,853	57,541	231	15,057	32,627	142
	2	27,596	56,910	387	15,134	32,861	238
	3	25,893	53,202	560	17,017	35,187	404
	4	20,319	38,710	828	22,590	42,073	903
Heart failure (hosp)	1	27,530	56,886	364	14,847	32,183	317
	2	27,486	56,644	582	15,183	32,622	482
	3	25,482	52,475	965	16,500	34,234	865
	4	19,704	37,842	1,578	20,746	39,414	2,109
Stroke (ischemic or hemorrhagic)	1	27,291	56,413	375	14,734	31,988	229
	2	27,054	55,846	490	15,011	32,220	371
	3	24,763	50,976	752	16,209	33,763	629

4	18,666	35,775	979	20,905	39,444	1,094

3.3 Step 3: Prediction

We internally developed separate prediction models for acute MI in all 3 databases.

The prediction models were estimated on the propensity score matched (1:1) subset of the population, using caliper of 0.2 and after excluding patients having the outcomes any time prior to treatment initiation. We chose a time horizon of 2 years after inclusion into the target cohort. For this demonstration, we developed the prediction models using LASSO logistic regression with 3-fold cross validation for hyper-parameter selection. For this demonstration, we developed the prediction models using LASSO logistic regression with 3-fold cross validation for hyper-parameter selection.

The models developed in the 3 databases had moderate discriminative performance (internally validated) with no major issues of overfitting to any cohort except for the case of CCAE database in which the derived prediction model performed better in the comparator cohort (Table 2). We also observed lower performance of the prediction model developed in MDCR compared to the other 2 databases. Results on the calibration of the prediction models can be found in the supplement.

Table 2: Discriminiative ability (c-statistic) of the derived prediction models for acute myocardial infarction in the matched set (development set), the treatment cohort, the comparator cohort and the etnire population in CCAE, MDCD and MDCR

Population	CCAE	MDCD	MDCR
Matched set	75.78	78.12	71.09
Treatment cohort	70.65	78.29	69.03
Comparator cohort	79.10	80.26	68.53
Entire population	74.06	79.46	69.07

3.4 Step 4: Estimation

Our aim was to estimate the average treatment effects on the relative and the absolute scale within strata of predicted acute MI risk.

We used patient-level predictions to stratify the patient population into 4 risk quarters. Within risk strata, relative effects were estimated using Cox regression and absolute effects were estimated from the Kaplan-Meier estimate

differences at 2 years after treatment initiation. To adjust for observed confounding within risk strata, we estimated propensity scores using the same approach as in the development of prediction models. We used the estimated propensity scores to stratify patients into 5 strata, within each risk quarter.

In general, there was sufficient overlap of propensity score distribution in all risk strata (Figure 2). If empirical equipoise was not achieved, the validity of the comparative effectiveness estimates would be questionable.

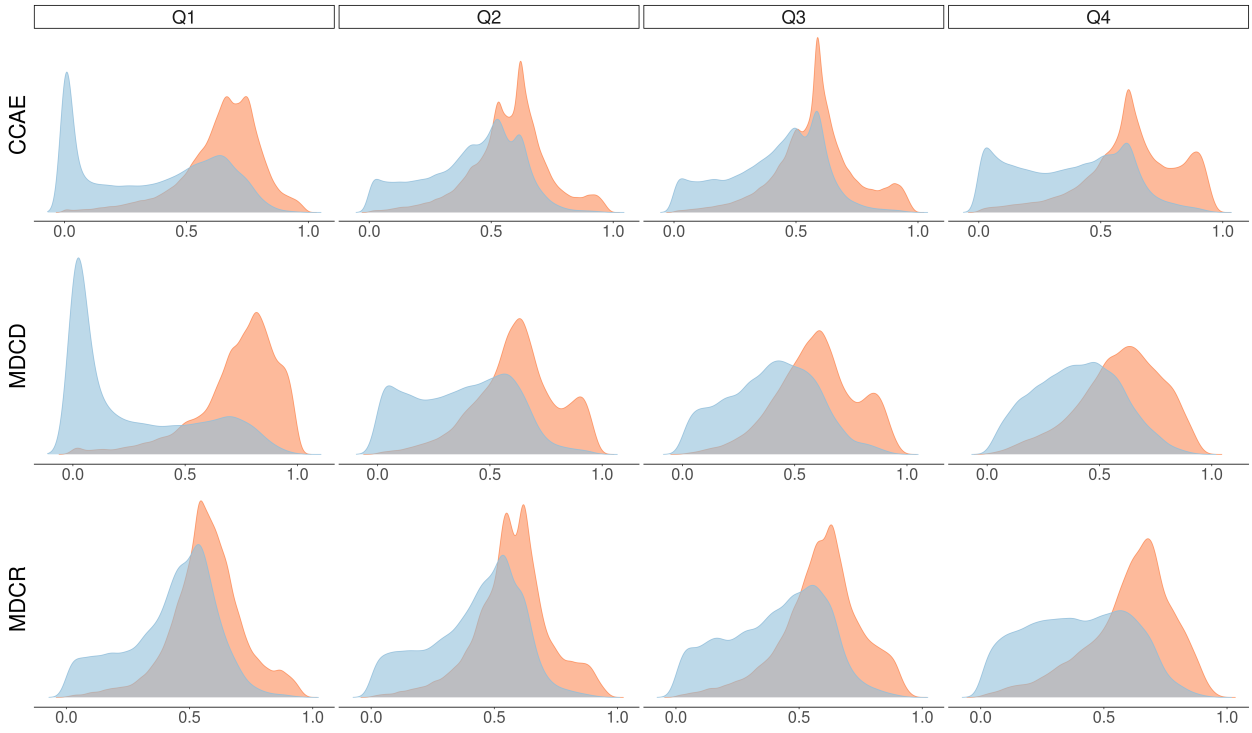

Figure 2: Preference score distributions for the evaluation of heterogeneity of the effect of ACE inhibitors compared to beta blockers on acute MI based on quarters of predicted acute MI risk. The preference score is a transformation of the propensity score that adjusts for prevalence differences between populations.

Propensity score adjustment achieved balance for most of the considered covariates, measured using standardized mean differences before and after adjustment (Figure 3). However, in lower risk strata imbalances persisted for a substantial subset of the covariates, all related to pregnancy findings. This was anticipated (use of ACE inhibitors is specifically contraindicated during pregnancy) and was already pointed out in [25].

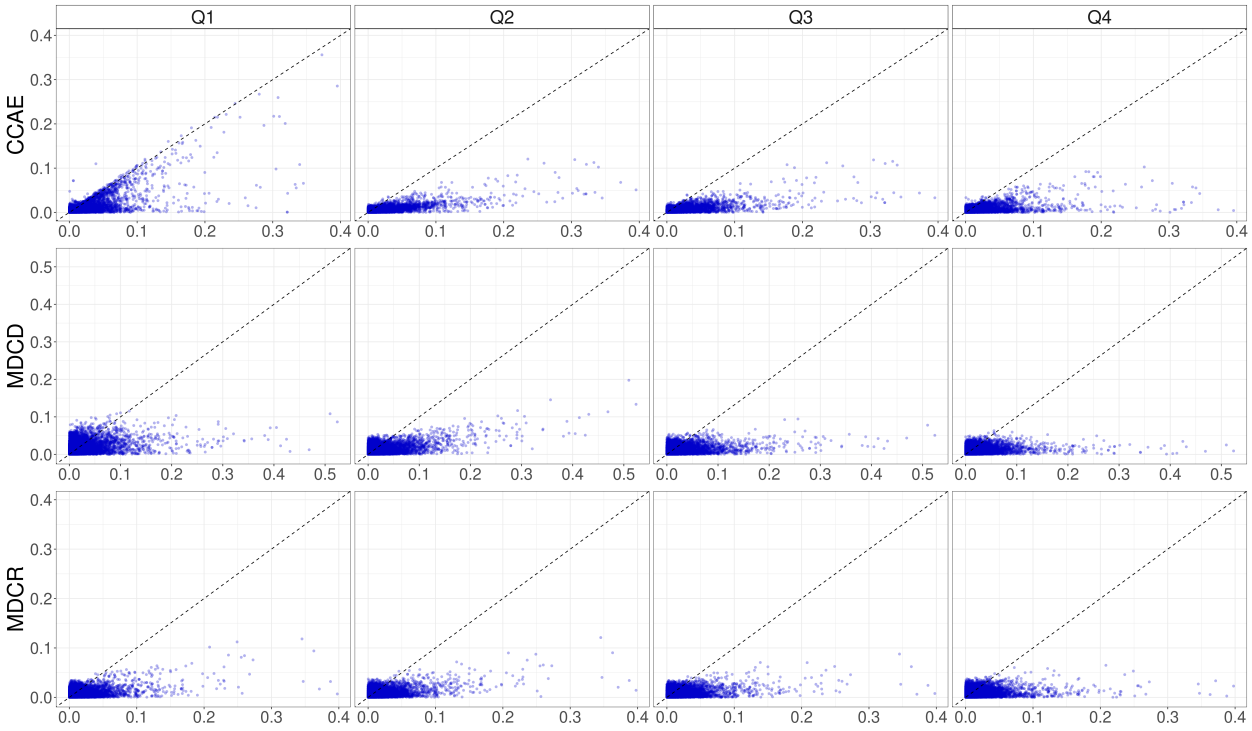

Figure 3: Patient characteristic balance for ACE inhibitors and beta blockers before and after stratification on the propensity scores. Each dot represents the standardized difference of means for a single covariate before (x-axis) and after (y-axis) stratification

3.5 Step 5: Presentation of results

For hospitalization with acute MI there was an increasing trend in favor ACE inhibitors compared to beta blockers on the relative scale (hazard ratios decreased) with increasing acute MI risk. More specifically, hazard ratios decreased from 0.98 (0.77 to 1.26; 95% CI), 1.30 (0.51 to 3.22; 95% CI) and 1.03 (0.82 to 1.29; 95% CI) to 0.76 (0.71 to 0.82; 95% CI), 0.94 (0.82 to 1.07; 95% CI) and 1.03 (0.93 to 1.15; 95% CI) in CCAE, MDCD and MDCR respectively (Figure 4). In terms of hospitalization with heart failure relative treatment effect estimates favored ACE inhibitors across all risk strata in all databases. We found no differences between the two treatments in their effect on stroke on the relative scale. In terms of the safety outcomes we found an increased risk of cough and angioedema on the relative scale across all risk strata. In the case of cough, this effect decreased with increasing risk of acute MI—from 1.37 (1.33 to 1.41; 95% CI), 1.35 (1.24 to 1.48; 95% CI) and 1.37 (1.29 to 1.45; 95% CI) in the lowest acute MI risk quarter to 1.26 (1.22 to 1.29; 95% CI), 1.07 (1.00 to 1.14) and 1.10 (1.04 to 1.17; 95% CI) in the highest acute MI risk quarter in CCAE, MDCD and MDCR respectively.

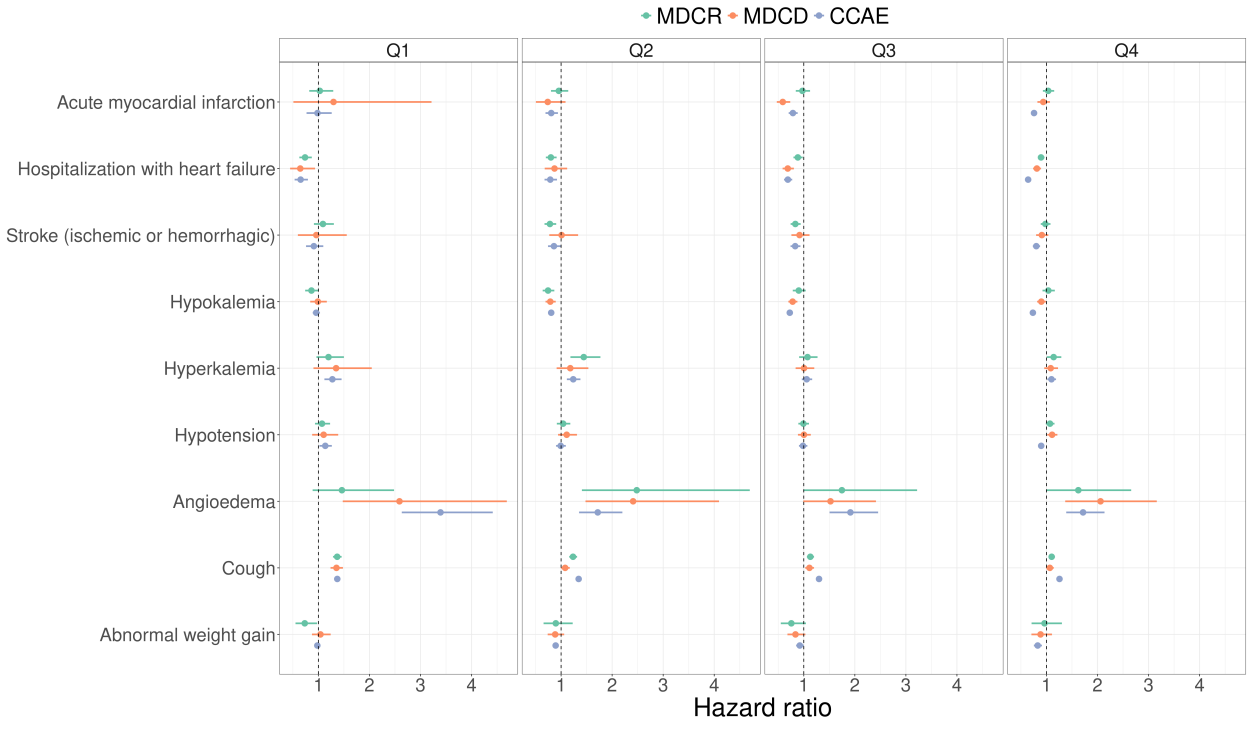

Figure 4: Overview of heterogeneity of ACE-inhibitors treatment on the relative scale (hazard ratios) within strata of predicted risk of acute MI. Values below 1 favor ACE inhibitors, while values above 1 favor beta blockers.

We observed an increasing trend of treatment effect on the absolute scale with increasing acute MI risk in favor of

ACE inhibitors in terms of acute MI in all databases except for MDCR—from -0.03% (-0.08% to 0.01%; 95% CI), -0.05% (-0.18% to 0.08%; 95% CI) and -0.02% (-0.24% to 0.19%; 95% CI) in the lowest acute MI risk quarter to 0.54% (0.36% to 0.71%; 95% CI), 0.29% (-0.39% to 0.97%; 95% CI) and -0.39% (-0.96% to 0.18%; 95% CI) in the highest acute MI risk quarter in CCAE, MDCD and MDCR, respectively (Figure 5). We found no difference on the absolute scale for stroke across risk strata. Absolute risk differences did not favor ACE inhibitors compared to beta blockers in terms of cough, even though this effect again diminished with increasing acute MI risk—from -4.14% (-4.62% to -3.66%; 95% CI), -6.45% (-9.12% to -3.78%; 95% CI) and -4.81% (-5.76% to -3.85%; 95% CI) in the lowest acute MI risk quarter to -2.57% (-2.99% to -2.15%; 95% CI), -1.11% (-2.93% to 0.70%; 95% CI) and -1.69% (-2.83% to -0.55%; 95% CI) in the highest acute MI risk quarter in CCAE, MDCD and MDCR, respectively. In terms of angioedema absolute risk differences were very small due to the rarity of the outcome.

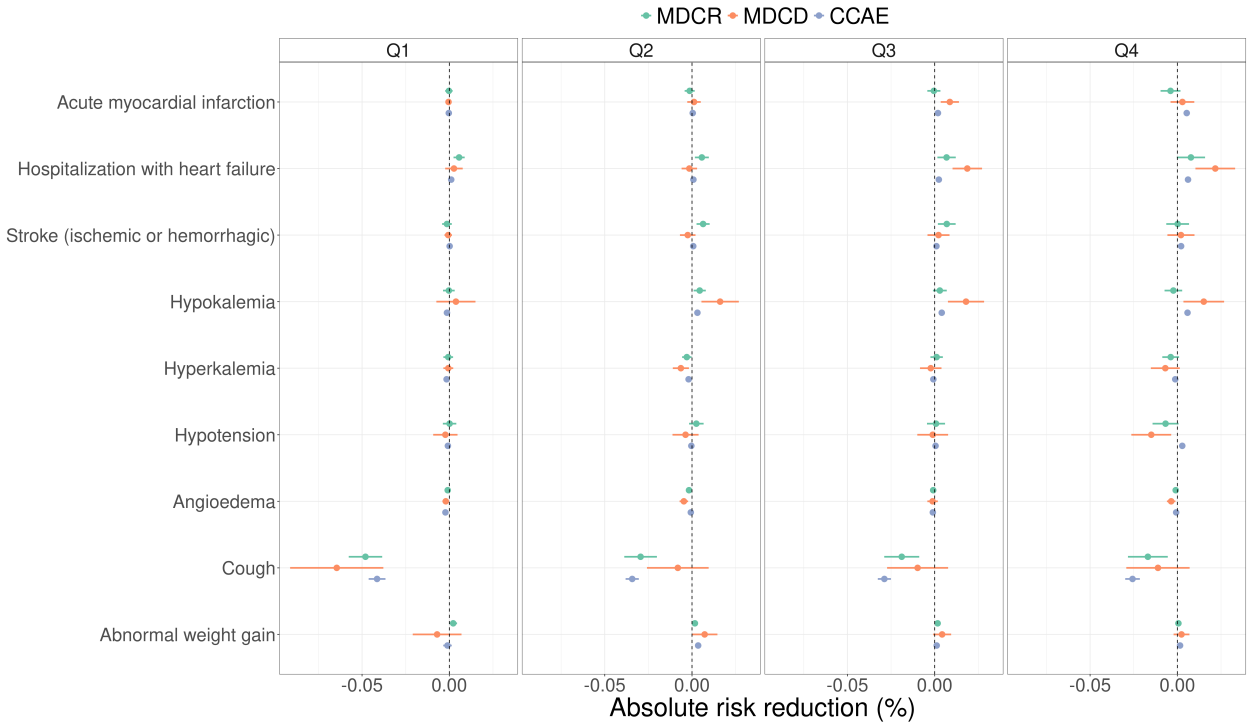

Figure 5: Overview of heterogeneity of ACE-inhibitors treatment on the absolute scale within strata of predicted risk of acute MI. Estimates of absolute treatment effect are derived as the difference in Kaplan-Meier estimates at 730 after inclusion. Values above 0 favor ACE inhibitors, while values below 0 favor beta blockers.

These results suggest that treatment with ACE-inhibitors, compared to treatment with beta blockers, may be focused on the higher risk patients, in whom the benefits outweigh the harms, while beta blockers may be a viable option in lower risk patients, in whom the benefit-harm tradeoff is more favorable. This is in accordance with

earlier findings that beta blockers should also be considered as first-line treatment for younger hypertensive patients [23,26]. This analysis, however, was carried out as demonstration of the framework and more rigorous analyses are required to make any suggestions for clinical practice.

The results of the analyses performed can be accessed and assessed through a publicly available web application (https://data.ohdsi.org/AceBeta9Outcomes).

4 Discussion

We developed a framework for the assessment of heterogeneity of treatment effect in large observational databases using a risk modeling approach. The framework is implemented in an open source R-package in the OHDSI methods library (https://github.com/OHDSI/RiskStratifiedEstimation). As a proof-of-concept, we used our framework to evaluate heterogeneity of the effect of treatment with ACE-inhibitors compared to beta blockers on 3 efficacy and 6 safety outcomes.

In recent years several methods for the evaluation of treatment effect heterogeneity have been developed in the setting of RCTs [5]. 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. Our framework provides a standardized approach that can be used to leverage available information from these data sources, allowing for large-scale risk-based assessment of treatment effect heterogeneity. It is an addition to the rapidly expanding literature of approaches for evaluating treatment effect heterogeneity. Multiple outcomes can be evaluated in patient subgroups of similar baseline outcome risk. Multiple outcome risk stratification schemes can also be considered. However, this should be done with caution, as it may hinder the interpretability of the results, in a similar manner as typical subgroup analyses.

Recently, guidelines on the application of risk modeling approaches for the assessment of heterogeneity of treatment effect in RCT settings have been proposed [27,28]. Our framework aims to translate these guidelines to the observational setting while also providing a toolkit for its implementation within OHDSI. It encourages open science as it requires accurate definition of the research questions translated into clear and reproducible cohort definitions that can easily be shared among researchers. Researchers with access to different databases mapped to OMOP-CDM can also very easily extend their overall analyses with risk-based assessment of treatment effect heterogeneity. This enables collaboration among multiple sites with access to different patient populations. We

propose that the framework is implemented any time treatment effect estimation in high-dimensional observational data is undertaken.

Several considerations need to be made. First, estimates may be biased due to the observational nature of the data. We attempt to account for potential confounding by estimating propensity scores within strata of predicted risk. These scores are estimated using regularized logistic regression on a large set of pre-defined covariates. However, such approaches do not account for unobserved confounding [29]. Several sensitivity analyses have been proposed in the literature for measuring the robustness of results in the presence of unobserved confounding. Another approach is to calibrate estimates and confidence intervals based on a large set of negative controls [30,31]. Negative controls are treatment-outcome pairs for which a null effect has been established. Estimating these effects within available data provides an approximation of the null distribution that can be used to empirically recalibrate effect estimates. Future work may extend our framework with this type of analyses.

Our method provides a risk-stratified assessment of treatment effect heterogeneity. However, even though stratification can provide a useful overview for clinical interpretation, these results cannot be applied to individuals in a straightforward manner, as we are still estimating subgroup effects [27]. Presentation of treatment effects as a continuous function of risk would be more helpful, but is methodologically challenging. Future research is necessary for the development of methods for continuous risk-based assessment of HTE.

Ideally, externally derived and adequately validated prediction model would be preferred for analyzing treatment effect heterogeneity [6]. In the absence of such prediction models an internally-developed risk prediction model can be considered. Earlier simulations of RCT studies have shown that internal models developed on the combined treatment and control arms blinded to treatment gave relatively unbiased estimates of treatment effect across the spectrum of risk [16]. However, in observational databases treatment arms may significantly differ in sample size. Because the prediction model will possibly better fit to the larger treatment arm, spurious treatment-covariate interactions may be introduced in the prediction model, leading to sub-optimal risk stratification. As a remedy, we first match the patients in the treatment and the comparator cohorts on the basis of propensity scores. Additionally, we propose to assess model performance in the separate treatment arms to evaluate its aptness for risk stratification.

Recently, disease risk scores have been explored as an alternative to propensity scores for balancing covariates [32,33]. In our method, the objective of risk stratification is not balancing, but assessing the variation of treatment effects on multiple outcomes across patients with different levels of baseline risk. Although using the same risk model for balancing and risk-based HTE analysis may sound attractive, we note that our method only uses one risk model for stratification and one propensity score model for balancing, while separate disease risk score models would be required to analyze treatment effects for each of the multiple outcomes.

In conclusion, the proof-of-concept study demonstrates the feasibility of our framework for risk-based assessment of treatment effect heterogeneity in large observational data. The standardized framework is easily applicable and highly informative whenever treatment effect estimation in high-dimensional observational data is of interest. Our framework is a supplement to the population-level effect estimation framework developed within OHDSI and, in the presence of an adequately discriminating prediction model, can be used to make the overall results more actionable for medical decision making.

5 References

1 Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly* 2004;**82**:661–87. doi:10.1111/j.0887-378x.2004.00327.x

2 Yusuf S. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA: The Journal of the American Medical Association* 1991;**266**:93. doi:10.1001/jama.1991.03470010097038

3 Garcia EF y, Nguyen H, Duan N *et al.* Assessing heterogeneity of treatment effects: Are authors misinterpreting their results? *Health Services Research* 2010;**45**:283–301. doi:10.1111/j.1475-6773.2009.01064.x

4 Varadhan R, Segal JB, Boyd CM *et al.* A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology* 2013;**66**:818–25. doi:10.1016/j.jclinepi.2013.02.009

5 Rekkas A, Paulus JK, Raman G *et al.* Predictive approaches to heterogeneous treatment effects: A scoping review. *BMC Medical Research Methodology* 2020;**20**. doi:10.1186/s12874-020-01145-1

6 Kent DM, Rothwell PM, Ioannidis JP *et al.* Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials* 2010;**11**. doi:10.1186/1745-6215-11-85

7 Kent DM, Steyerberg E, Klaveren D van. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *BMJ* 2018;k4245. doi:10.1136/bmj.k4245

8 Adler-Milstein J, Holmgren AJ, Kralovec P *et al.* Electronic health record adoption in US hospitals: The emergence of a digital "advanced use" divide. *Journal of the American Medical Informatics Association* 2017;**24**:1142–8. doi:10.1093/jamia/ocx080

9 Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? *JAMA* 2014;**312**:129. doi:10.1001/jama.2014.4364

10 Hripcsak G, Duke JD, Shah NH *et al.* Observational health data sciences and informatics (ohdsi): Opportunities for observational researchers. *Studies in health technology and informatics* 2015;**216**:574.

11 Overhage JM, Ryan PB, Reich CG *et al.* Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* 2012;**19**:54–60. doi:10.1136/amiajnl-2011-000376

12 Reps JM, Schuemie MJ, Suchard MA *et al.* 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* 2018;**25**:969–75. doi:10.1093/jamia/ocy032

13 Ryan PB, Schuemie MJ, Gruber S *et al.* Empirical performance of a new user cohort method: Lessons for developing a risk identification and analysis system. *Drug Safety* 2013;**36**:59–72. doi:10.1007/s40264-013-0099-6

14 Collins GS, Reitsma JB, Altman DG *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMC Medicine* 2015;**13**:1. doi:10.1186/s12916-014-0241-z

15 Moons KG, Altman DG, Reitsma JB *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine* 2015;**162**:W1. doi:10.7326/m14-0698

16 Burke JF, Hayward RA, Nelson JP *et al.* Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circulation: Cardiovascular Quality and Outcomes* 2014;**7**:163–9. doi:10.1161/circoutcomes.113.000497

17 Klaveren D van, Balan TA, Steyerberg EW *et al.* Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology* 2019;**114**:72–83. doi:10.1016/j.jclinepi.2019.05.029

18 Kent DM, Nelson J, Dahabreh IJ *et al.* Risk and treatment effect heterogeneity: Re-analysis of individual participant data from 32 large clinical trials. *International Journal of Epidemiology* 2016;dyw118. doi:10.1093/ije/dyw118

19 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011;**46**:399–424. doi:10.1080/00273171.2011.568786

20 Wiysonge CS, Bradley HA, Volmink J *et al.* Beta-blockers for hypertension. *Cochrane database of systematic reviews* 2017.

21 Whelton PK, Carey RM, Aronow WS *et al.* 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* 2018;**71**. doi:10.1161/hyp.0000000000000065

22 Williams B, Mancia G, Spiering W *et al.* 2018 ESC/ESH guidelines for the management of arterial hypertension. *European Heart Journal* 2018;**39**:3021–104. doi:10.1093/eurheartj/ehy339

23 Khan N. Re-examining the efficacy of -blockers for the treatment of hypertension: A meta-analysis. *Canadian Medical Association Journal* 2006;**174**:1737–42. doi:10.1503/cmaj.060110

24 Rabi DM, McBrien KA, Sapir-Pichhadze R *et al.* Hypertension canada's 2020 comprehensive guidelines for the prevention, diagnosis, risk assessment, and treatment of hypertension in adults and children. *Canadian Journal of Cardiology* 2020;**36**:596–624. doi:10.1016/j.cjca.2020.02.086

25 Suchard MA, Schuemie MJ, Krumholz HM *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis. *The Lancet* 2019;**394**:1816–26. doi:10.1016/s0140-6736(19)32317-7

26 Cruickshank J. Are we misunderstanding beta-blockers. *International Journal of Cardiology* 2007;**120**:10–27. doi:10.1016/j.ijcard.2007.01.069

27 Kent DM, Paulus JK, Klaveren D van *et al.* The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine* 2019;**172**:35. doi:10.7326/m18-3667

28 Kent DM, Klaveren D van, Paulus JK *et al.* The predictive approaches to treatment effect heterogeneity (path) statement: Explanation and elaboration. *Annals of Internal Medicine* 2020;**172**:W1–W25. doi:10.7326/M18-3668

29 Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science* 2013;**14**:570–80. doi:10.1007/s11121-012-0339-5

30 Schuemie MJ, Ryan PB, DuMouchel W *et al.* Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine* 2014;**33**:209–18. doi:https://doi.org/10.1002/sim.5925

31 Schuemie MJ, Hripcsak G, Ryan PB *et al.* Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences* 2018;**115**:2571–7. doi:10.1073/pnas.1708282114

32 Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiology and Drug Safety* 2012;**21**:138–47. doi:10.1002/pds.3231

33 Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008;**95**:481–8. doi:10.1093/biomet/asn004