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, MD[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

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 the RCT setting. The aim of this study was to extend this approach to the observational setting using a standardised 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 angiotensin-converting enzyme (ACE) inhibitors versus beta blockers on three efficacy and six safety outcomes across three observational databases. The proposed framework can supplement any comparative effectiveness study. 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 received negligible absolute benefits for all three efficacy outcomes, though they were more pronounced in the highest risk quarter, especially for hospitalisation with heart failure. However, failing diagnostics showed evidence of residual imbalances even after adjustment for observed confounding. 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. It is easily applicable and highly informative whenever treatment effect estimation in observational data is of interest.

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

INTRODUCTION

Treatment effects can often vary substantially across individual patients, causing overall effect estimates to be inaccurate for a significant proportion of the patients at hand[1,2]. Understanding heterogeneity of treatment effects (HTE) 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 HTE in clinical trials is through subgroup analyses, which are rarely adequately powered and can lead to false conclusions of absence of HTE or exaggerate its presence[3,4]. In addition, patients differ with regard to multiple characteristics simulatneously, resulting in much richer HTE compared to the one explored with regular on-variable-at-a-time subgroup analyses [Kent, BMJ 2018].

Baseline risk is a summary score inherently related to treatment effect that can represent more closely the variability in patient characteristics[3,5–8]. For example, an invasive coronary procedure—in comparison with medical treatment—improves survival in patients with myocardial infarction at high (predicted) baseline risk but not in those at low baseline risk[9]. It has also been shown that high-risk patients with pre-diabetes benefit substantially more from a lifestyle modification program than low-risk patients[10].

Recently, systematic guidance on the application of risk-based methods for the assessment of HTE has been developed for RCT data[11,12]. 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 HTE analysis 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[12].

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)[13,14]. A standardized framework applying current best practices for comparative effectiveness studies within the OHDSI setting has been proposed[15]. This framework was successfully implemented on a large scale for estimation of average effects of all first-line hypertension treatment classes on a total of 52 outcomes of interest across a global network of nine observational databases[16].

We aimed to develop a framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases, extending the existing methodology developed for the RCT setting. We implemented the framework in a publicly available package providing an out-of-the-box solution for implementing such analyses at scale within any observational database mapped to OMOP-CDM. In a case study we analyzed heterogeneity of the effects of first-line hypertension treatment.

RESULTS

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.

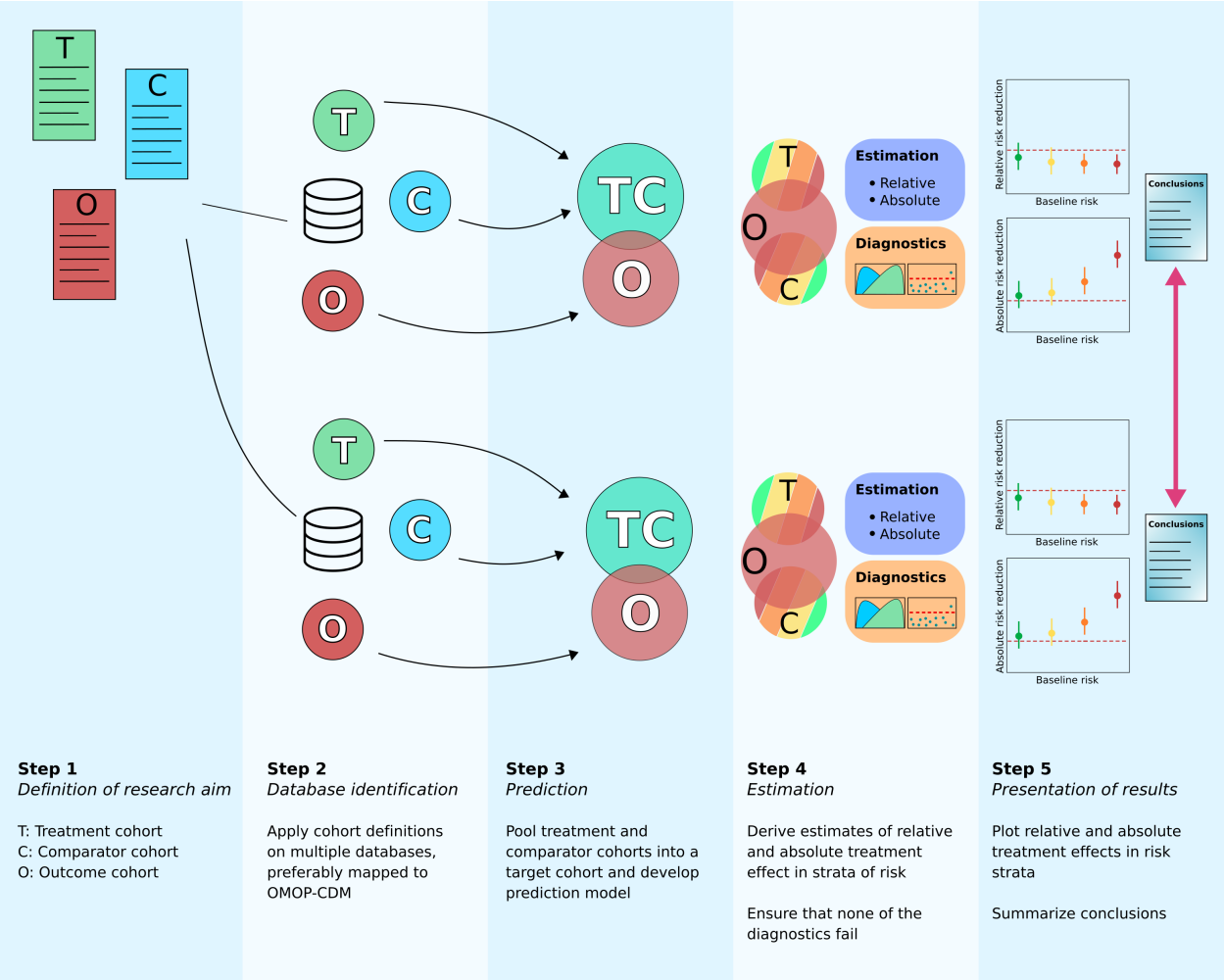

Figure 1: Framework execution diagram. Illustration of how the framework is applied on two observational databases, preferably mapped to OMOP-CDM.

As a demonstration, we evaluated if our proposed method was able to identify treatment effect heterogeneity of ACE inhibitors compared to beta blockers using acute myocardial infarction (MI) risk quarter specific effect estimates, both on the relative and on the absolute scale. We focused on three efficacy outcomes (acute MI,

hospitalization with heart failure and ischemic or hemorrhagic stroke) and six safety outcomes (hypokalemia, hyperkalemia, hypotension, angioedema, cough and abnormal weight gain). 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 ACE inhibitors (T) to the effect of beta blockers (C) in patients with established hypertension with respect to nine 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 considered three efficacy and six 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); hypokalemia; hyperkalemia; hypotension; angioedema; cough; abnormal weight gain (safety outcomes). Among the safety outcomes we focus on angioedema and cough which are two known adverse events linked to treatment with ACE inhibitors[17]. Results on the rest of the safety outcomes are included in the supplement.

All cohort definitions were identical to the ones used in the multinational study that provided overall treatment effect estimates comparing all anti-hypertensive drug classes with each other[16]. More information can be found in the supplementary material.

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 Medicaid (MDCD), and IBM MarketScan Medicare Supplemental Beneficiaries (MDCR). Our analyses included a total of 924,459, 107,046, and 106,905 patients initiating treatment with ACE inhibitors and 465,763, 76,546, and 73,213 patients initiating treatment with beta blockers in CCAE, MDCD and MDCR respectively (Table 1). Adequate numbers of patients were included in all strata of predicted acute MI risk (Supplementary Tables S1-S3).

Table 1: Number of patients, person years and events within quarters of predicted risk for acute MI for the three efficacy outcomes of the study (acute MI, hospitalization with heart failure and ischemic or hemorrhagic stroke) across the three databases.

	ACE inhibitors			Beta blockers		
Outcome	Patients	Person years	Outcomes	Patients	Person years	Outcomes
CCAE						
acute myocardial infarction	924,196	1,327,973	4,102	457,375	648,612	2,492
hospitalization with heart failure	924,459	1,328,430	3,764	465,763	660,580	3,711
stroke	917,501	1,319,236	3,741	464,989	659,472	2,454
MDCD						
acute myocardial infarction	107,046	162,590	1,448	76,307	112,767	1,361
hospitalization with heart failure	105,544	160,237	2,819	74,649	110,455	3,005
stroke	104,953	159,344	1,799	76,546	113,048	1,623
MDCR						
acute myocardial infarction	106,905	163,260	1,764	72,733	110,821	1,480
hospitalization with heart failure	106,191	162,258	3,004	73,182	111,710	3,592
stroke	103,531	158,369	2,323	73,213	111,613	2,241

Step 3: Prediction

We internally developed separate prediction models for acute MI in all three databases. The prediction models were estimated on the propensity score matched (1:1) subset of the sample, using caliper of 0.2 and after excluding patients having the outcome any time prior to treatment initiation. We chose a 2-year time at risk for patients and developed the prediction models using LASSO logistic regression with 3-fold cross validation for hyper-parameter selection.

The models had moderate discriminative performance (internally validated) with no major issues of overfitting to any cohort except for the case of CCAE, where 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.

Table 2: Discriminative 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 entire population in CCAE, MDCD and MDCR.

Population	CCAE	MDCD	MDCR
Matched	0.73 (0.72, 0.74)	0.78 (0.77, 0.79)	0.66 (0.65, 0.68)
Treatment	0.69 (0.68, 0.70)	0.76 (0.75, 0.77)	0.65 (0.63, 0.66)
Comparator	0.77 (0.76, 0.78)	0.82 (0.81, 0.82)	0.64 (0.63, 0.66)

Entire population	0.72 (0.71, 0.73)	0.79 (0.78, 0.80)	0.65 (0.64, 0.66)

Step 4: Estimation

We used patient-level predictions to stratify the sample into four acute MI risk quarters. Within risk quarters, relative effects were estimated using Cox regression and absolute effects were derived from the Kaplan-Meier estimate differences at two years after treatment initiation. To adjust for observed confounding within each risk quarter, we estimated propensity scores using the same approach as step 3 and stratified patients into five propensity score strata. The risk quarter-specific effect estimates were derived by averaging over the estimates within the propensity score fifths.

In the lowest acute MI risk quarter of CCAE and MDCD we observed strong separation of the propensity score distributions, therefore, effect estimates derived in these strata are not well-supported (Figure 2). This problematic behavior is also visible in the covariate balance plots comparing standardized mean differences of patient characteristics before and afrer PS adjustment, where in many cases the commonly accepted bound of 0.1 is violated (Figure 3). This is more pronounced in the lowest acute MI risk quarter of CCAE, but remains an issue for a small number of covariates in all CCAE risk strata. This diagnostic also fails for the two lower acute MI risk quarters of MDCD. Often the persisting imbalances were linked to pregnancy outcomes, which can be explained by the contraindication of ACE inhibitors in this condition. Analyses in MDCR passed all diagnostics.

Finally, the distribution of the estimated relative risks with regard to 30 negative control outcomes indicated unresolved confounding within the lowest acute MI risk quarter of CCAE (Figrue 4). Hazard ratios significantly different than 1 (true effect size) were concentrated in the lower right part of Figure 4: panel Q1. This suggests significant negative effects of ACE inhibitors compared to beta blockers on causally unrelated outcomes, pointing at unresolved differences between the two treatment arms. This was not the case in the other risk quarters of CCAE, or in any risk quarter of MDCD and MDCR (Supplementary Figures S1-S2).

Step 5: Presentation of results

The overall estimated hazard ratios for the main outcomes are presented in Table 3. 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 1.29 (1.00 to 1.68; 95% CI) and 1.58 (0.78 to 3.28; 95% CI) to 0.77 (0.71 to 0.83; 95% CI), 0.84 (0.76 to 0.94; 95% CI) in CCAE and MDCD respectively (Figure 5). In MDCR hazard ratios increased from 0.93 (0.75 to 1.17; 95% CI) in the

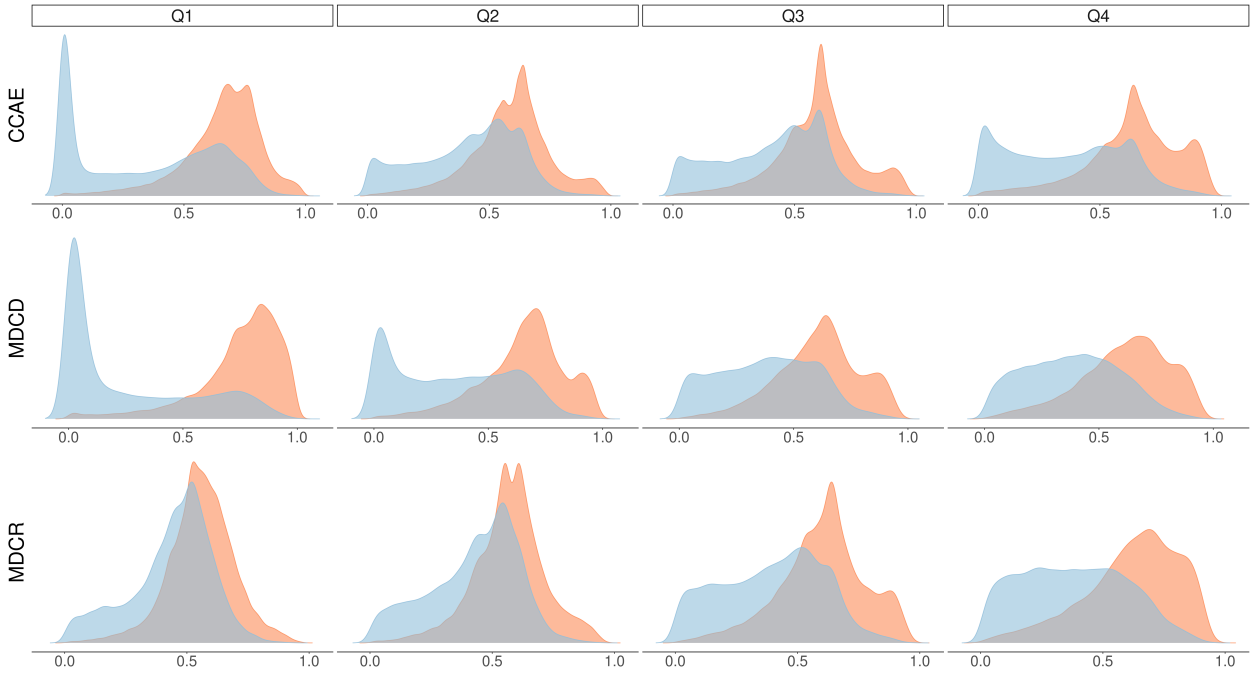

Figure 2: Preference score distributions in quarters of predicted acute MI risk. The preference score is a transformation of the propensity score that adjusts for prevalence differences between populations. Higher overlap of the preference score distributions indicates that patients in the target and the comparator cohorts are more similar in terms of the predicted probability of receiving treatment (ACE inhibitors).

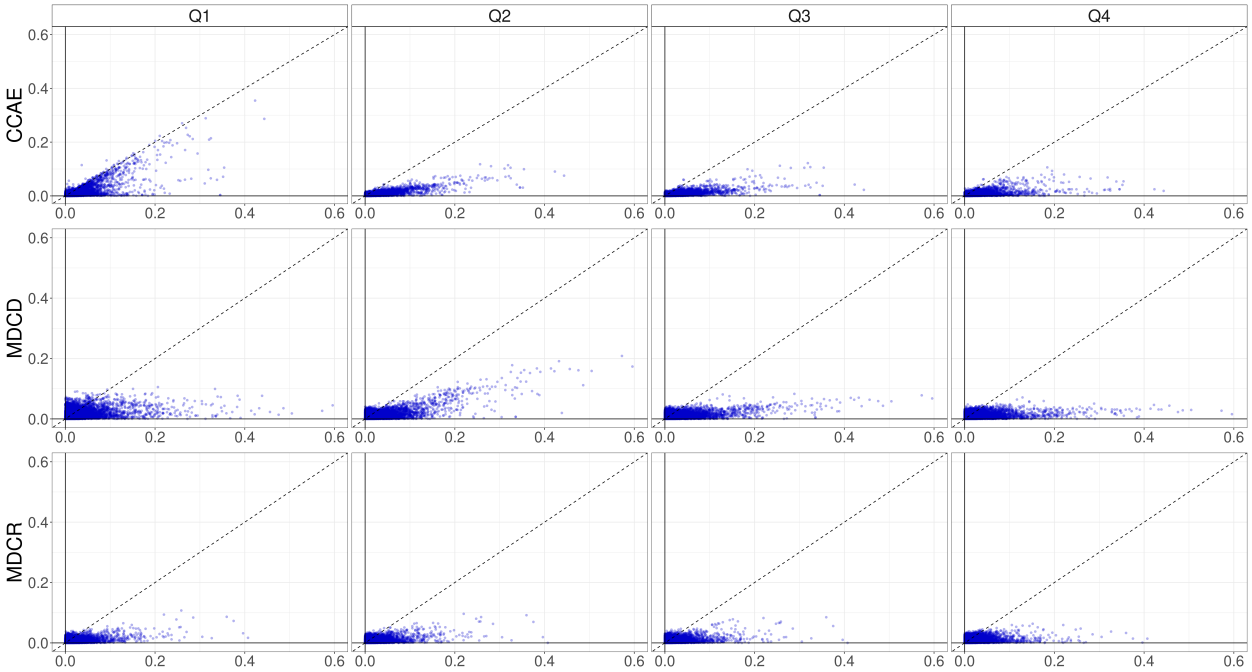

Figure 3: Patient characteristic balance for ACE inhibitors and beta blockers before and after stratification on the propensity scores. Each dot represents the standardized difference of means for a single covariate before (x-axis) and after (y-axis) stratification. A commonly used rule of thumb suggests that standardized mean differences above 0.1 indicate insufficient covariate balance post propensity score adjustment.

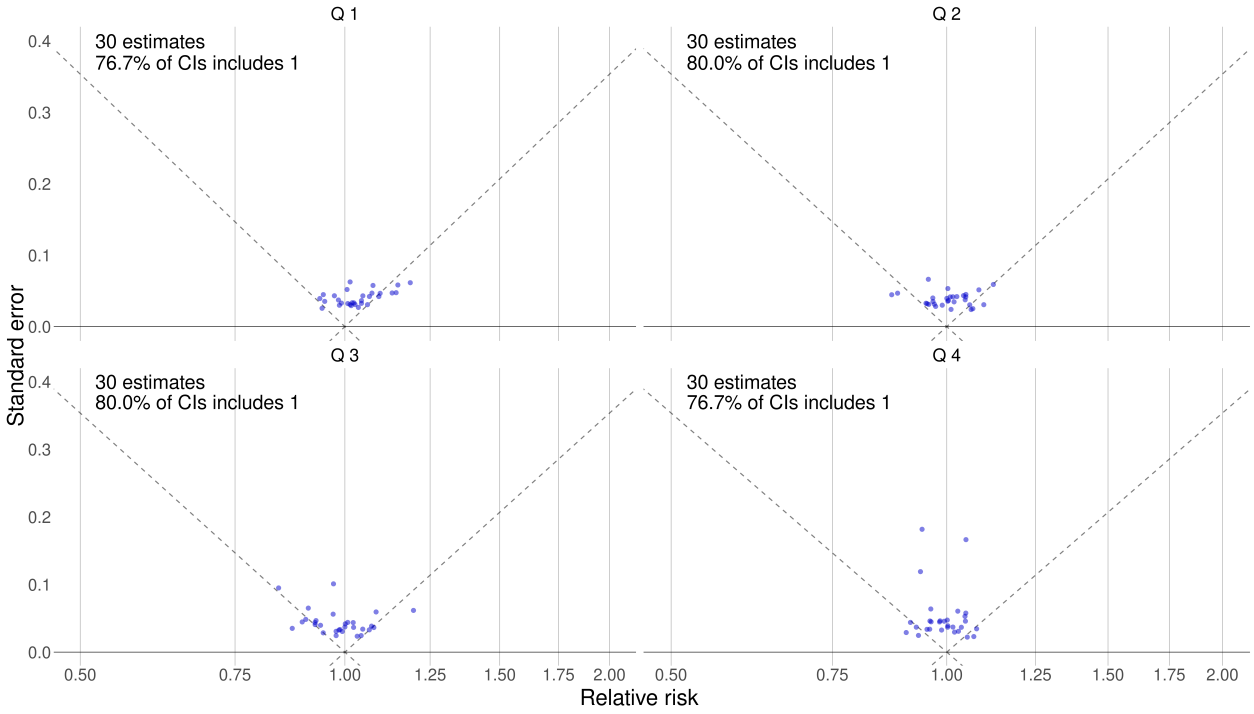

Figure 4: Systematic error. Effect size estimates for the negative controls (true hazard ratio = 1). Estimates below the diagonal dashed lines are statistically significant (alpha = 0.05) different from the true effect size. A well-calibrated estimator should have the true effect size within the 95 percent confidence interval 95 percent of times.

lowest MI risk quarter to 1.03 (0.92 to 1.16; 95% CI). Relative treatment effect estimates for hospitalization with heart failure favored ACE inhibitors across all risk strata in all databases. In the case of stroke in CCAE we found quite constant hazard ratios which became weaker in the highest risk quarter patients (0.88 with 95% CI from 0.80 to 0.96). In the other two databases no significant relative treatment effects were observed for stroke. In terms of the safety outcomes, we found an increased ACE inhibitor 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 1.41 (1.37 to 1.46; 95% CI), 1.28 (1.18 to 1.38; 95% CI), and 1.38 (1.29 to 1.48; 95% CI) to 1.30 (1.26 to 1.34; 95% CI), 1.06 (1.00 to 1.12; 95% CI), and 1.11 (1.04 to 1.18; 95% CI) in CCAE, MDCD, and MDCR, respectively.

Table 3: Overall hazard ratios.

Outcome	CCAE	MDCD	MDCR
acute myocardial infarction	0.83 (0.79, 0.88)	0.87 (0.80, 0.96)	1.02 (0.94, 1.10)
hospitalization with heart failure	0.66 (0.62, 0.69)	0.86 (0.81, 0.92)	0.85 (0.80, 0.90)
stroke	0.88 (0.83, 0.93)	0.97 (0.90, 1.05)	0.91 (0.85, 0.97)

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 -0.05% (-0.10% to -0.005%; 95% CI), -0.04% (-0.14% to 0.05%; 95% CI), and 0.08% (-0.19% to 0.34%; 95% CI) in the lowest acute MI risk quarter to 0.47% (0.31% to 0.63%; 95% CI), 0.93% (0.35% to 1.50%; 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 6). 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 -3.97% (-4.40% to -3.54%; 95% CI), -4.54% (-6.97% to -2.12%; 95% CI), and -3.64% (-4.60% to -2.68%; 95% CI) in the lowest acute MI risk quarter to -2.57% (-3.02% to -2.13%; 95% CI), -0.20% (-1.58% to 1.17%; 95% CI), and -1.08% (-2.25% to 0.08%; 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.

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

Interpretation

The overall benefits of ACE inhibitors compared to beta blockers for acute MI and hospitalization with heart failure are driven mainly by the higher acute MI risk patients in CCAE and MDCD, hence the observed increasing

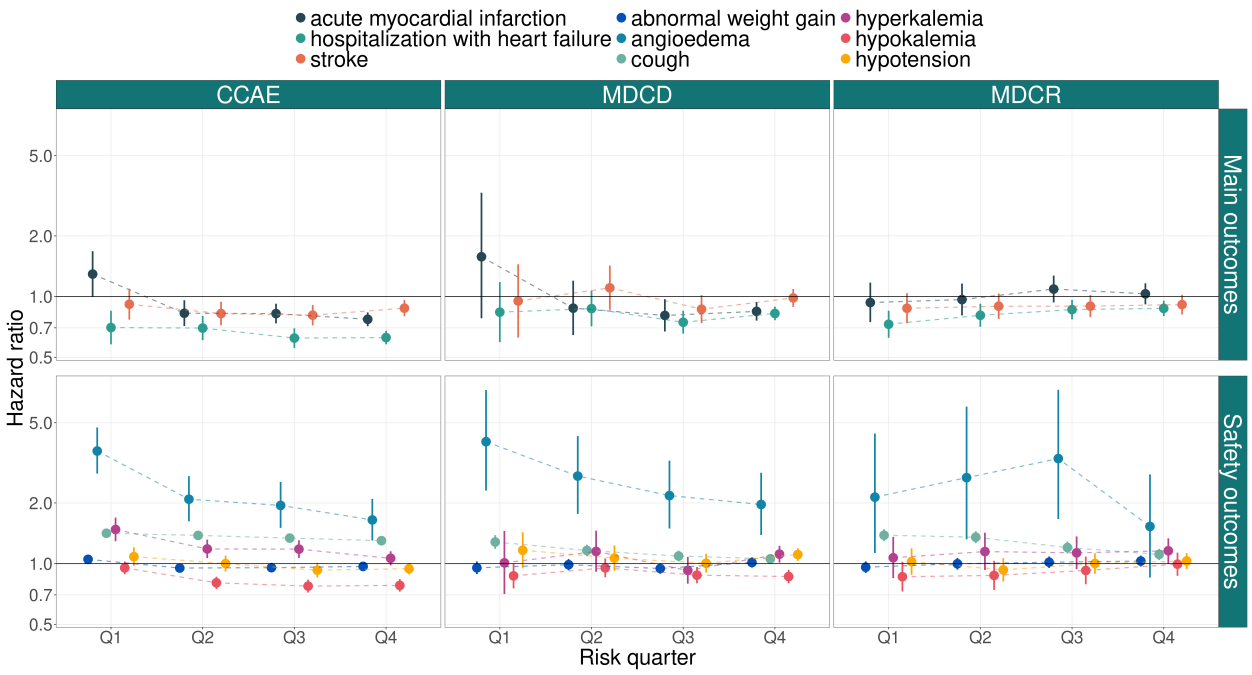

Figure 5: Overview of heterogeneity of ACE inhibitors treatment on the relative scale (hazard ratios) within strata of predicted risk of acute MI. Values below 1 favor ACE inhibitors, while values above 1 favor beta blockers.

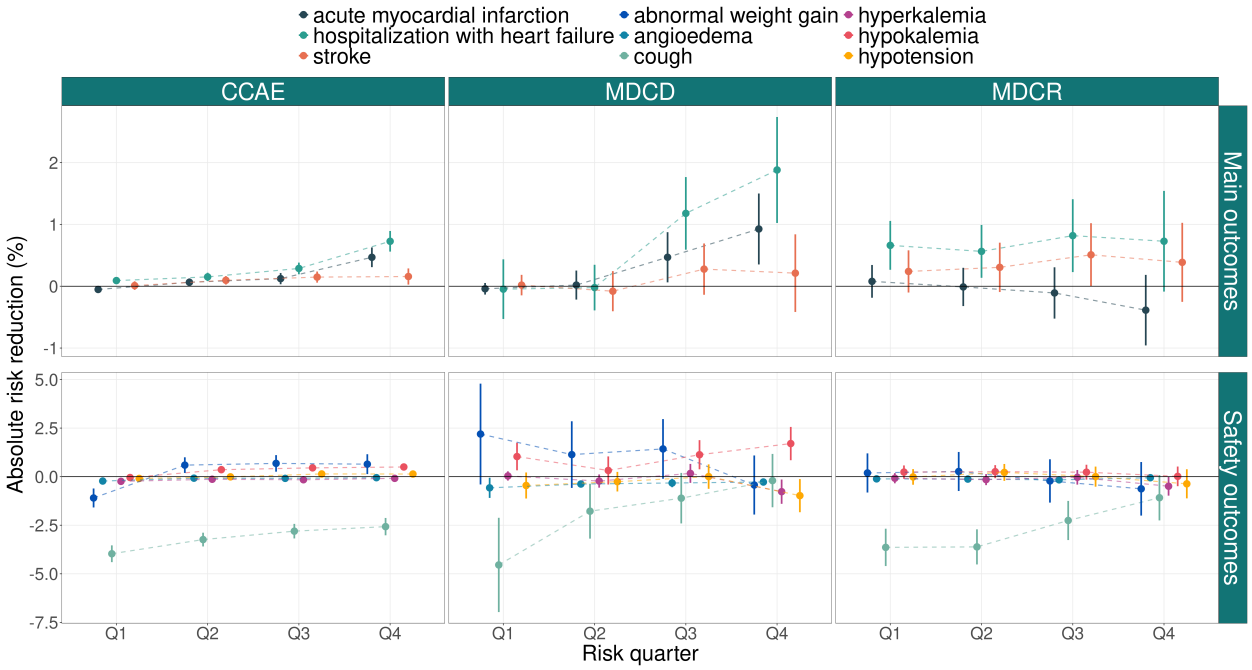

Figure 6: Overview of heterogeneity of ACE inhibitors treatment on the absolute scale within strata of predicted risk of acute MI. Estimates of absolute treatment effect are derived as the difference in Kaplan-Meier estimates at two years after inclusion. Values above 0 favor ACE inhibitors, while values below 0 favor beta blockers.

patterns of the absolute treatment benefits. In MDCR we found no significant overall difference on the relative scale for acute MI and, cosequently, no differences in acute MI risk strata were observed on any scale. For heart failure, MDCR patients at the lower half of acute MI risk had lower absolute benefits compared to the patients at the upper half. Finally, the small overall relative effect for stroke resulted in smaller absolute benefits of ACE inhibitors across acute MI risk strata in all databases.

For patients at lower acute MI risk, the cough and angioedema risk increase related to treatment with ACE inhibitors may be important factors to consider for medical decision making, given the small benefits observed for the main outcomes. However, diagnostics failed in lower risk patients within CCAE and MDCR which renders these conclusions less dependable.

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

DISCUSSION

The major contribution of our work is the development of a risk-based framework for the assessment of treatment effect heterogeneity in large observational databases. This fills a gap identified in the literature after the development of guidelines for performing such analyses in the RCT setting[11,12]. As an additional contribution we developed the software for implementing this framework in practice and made 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 demonstrated the use of our framework for the evaluation of treatment effect heterogeneity ACE inhibitors compared to beta blockers on three efficacy and six 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[18]. 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 subpopulations of interest for 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 HTE or falsely introduce it when there is none[19]. Therefore, inferences on both overall treatment effect estimates and HTE need to rely on strong, often unverifiable, assumptions, despite the advancements and guidance on best practices.

However, well-designed observational studies on average replicate RCT results, even though often differences in magnitude may occur[24]. 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[25].

Our framework highlights the scale dependency of HTE and how it relates to baseline risk. Treatment effect is mathematically determined by baseline risk, if we assume a constant non-zero effect size[26]. Patients with low baseline risk can only experience minimal benefits, before their risk is reduced to zero. In contrast, high risk patients are capable of displaying much higher absolute benefits. This becomes evident when evaluating the effects of ACE inhibitors on cough and angioedema, compared to treatment with beta blockers. 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. Conversely, in the case of angioedema, the substantial relative risk increase with ACE inhibitors only translated in a small absolute risk increase due to the quite low baseline angioedema risk.

The application of our framework in the case study is for demonstration purposes and there are several limitations to its conclusions. First, death could be a competing risk. We could expand our framework in the future to potentially support subdistribution hazard ratios and cumulative incidence reductions. Second, 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 others that were not assessed in our demonstration. Third, we did not to correct for multiplicity when presenting the results. We are interested in presenting trends in the data and not detecting the specific subgroups within which a non-null treatment effect is detected. 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 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.

Step 3: Prediction

Our method relies on adequately separating patients into subgroups based on their baseline risk for the outcomes of interest. Therefore, a model—either an existing external model adequately validated on an internally developed one—assigning patient-level risk is required. For internally developing a risk prediction model we adopt the standardized framework focused on observational data that ensures adherence to existing guidelines[27–29].

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[30,31]. 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 calibration 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[32].

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[25,33,34].

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.

DATA AVAILABILITY

The claims data are proprietary and are not publicly accessible due to restricted user agreement. Database descriptions are available in the Supplementary Material.

CODE AVAILABILITY

The source code for the R-package that implements our framework can be found at https://github.com/OHDSI/RiskStratifiedEstimation. The code for implementing the proof of concept study presented here is publicly available at https://github.com/mi-erasmusmc/AceBeta9Outcomes.

REFERENCES

1 Rothwell PM. Can overall results of clinical trials be applied to all patients? *The Lancet* 1995; **345**: 1616–9.

2 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.

3 Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Medical Research Methodology* 2006; **6**. DOI:10.1186/1471-2288-6-18.

4 Kent DM, Steyerberg E, Klaveren D van. Personalized evidence based medicine: Predictive approaches to heterogeneous treatment effects. *Bmj* 2018; : k4245.

5 ROTHWELL P, MEHTA Z, HOWARD S, GUTNIKOV S, WARLOW C. From subgroups to individuals: General principles and the example of carotid endarterectomy. *The Lancet* 2005; **365**: 256–65.

6 Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients. *JAMA* 2007; **298**: 1209.

7 Kent DM, Alsheikh-Ali A, Hayward RA. Competing risk and heterogeneity of treatment effect in clinical trials. *Trials* 2008; **9**. DOI:10.1186/1745-6215-9-30.

8 Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials* 2010; **11**. DOI:10.1186/1745-6215-11-85.

9 Thune JJ, Hoefsten DE, Lindholm MG, *et al.* Simple risk stratification at admission to identify patients with reduced mortality from primary angioplasty. *Circulation* 2005; **112**: 2017–21.

10 Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: Risk based reanalysis of diabetes prevention program. *BMJ* 2015; **350**: h454–4.

11 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.

12 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.

13 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.

14 Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association* 2012; **19**: 54–60.

15 Ryan PB, Schuemie MJ, Gruber S, Zorych I, Madigan D. Empirical performance of a new user cohort method: Lessons for developing a risk identification and analysis system. *Drug Safety* 2013; **36**: 59–72.

16 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.

17 Israili ZH. Cough and angioneurotic edema associated with angiotensin-converting enzyme inhibitor therapy. *Annals of Internal Medicine* 1992; **117**: 234.

18 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.

19 Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology* 2013; **66**: 818–25.

20 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine* 2000; **342**: 1887–92.

21 Ioannidis JPA. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001; **286**: 821.

22 Dahabreh IJ, Sheldrick RC, Paulus JK, *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* 2012; **33**: 1893–901.

23 Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clinical Pharmacology & Therapeutics* 2017; **102**: 924–33.

24 Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database of Systematic Reviews* 2014; **2014**. DOI:10.1002/14651858.mr000034.pub2.

25 Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2018; **376**: 20170356.

26 Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: Understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology* 2016; **45**: 2184–93.

27 Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. 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.

28 Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMC Medicine* 2015; **13**: 1.

29 Moons KGM, 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.

30 Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. *Circulation: Cardiovascular Quality and Outcomes* 2014; **7**: 163–9.

31 Klaveren D van, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. *Journal of Clinical Epidemiology* 2019; **114**: 72–83.

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

33 Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine* 2014; **33**: 209–18.

34 Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. 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.

ACKNOWLEDGEMENTS

AR and PRR have received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 80696. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

AUTHOR CONTRIBUTIONS

AR, PRR, and DVK conceptualised the study. AR, DVK, EWS, and DMK developed the methodology. PBR, and PRR acquired the data and AR analysed the data. AR developed the software and wrote drafted the manuscript, which was critically reviewed by DVK, PBR, EWS, DMK, and PRR. AR, PBR, and PRR had full access to the raw data. All authors read and approved the manuscript and had final responsibility for the decision to submit for publication.

COMPETING INTERESTS

AR and PRR work for a group that received unconditional research grants from Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, and Servier. None of these grants result in a conflict of interest for the content of this paper. PBR is an employee of Janssen R&D, subsidiary of Johnson & Johnson. DVK, DMK, EWS have nothing to declare.