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 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 3 efficacy and 6 safety outcomes across three observational databases. Patients at low risk of acute myocardial infarction (MI) received negligible absolute benefits for all 3 efficacy outcomes, though they were more pronounced in the highest risk quarter, especially for hospitalization with heart failure. The substantial risk increase of cough and angioedema with ACE inhibitors across all risk strata suggests that beta blockers provide a viable alternative to patients at low risk of acute MI. 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

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) [1,2]. A standardized framework applying current best practices for comparative effectiveness studies within the OHDSI setting has been proposed [3]. This framework was 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 9 observational databases [4].

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 [5,6]. Understanding heterogeneity of treatment effects (HTE) has been central to the agenda 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. While exploratory and confirmatory HTE analyses focus on generating and testing hypotheses on subgroup effects, predictive HTE analyses use all available patient data to predict the benefits or harms of treatment in individual patients [7,8].

Baseline risk has been recognized as a robust predictor of treatment effect, because of the inherent relationship between baseline risk and the absolute effects of treatment [9–13]. 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 [14]. It has also been shown that high-risk patients with pre-diabetes benefit substantially more from a lifestyle modification program than low-risk patients [15].

Recently, systematic guidance on the application of risk-based methods has been developed for RCT data [16,17]. 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. Several methods for predictive HTE analysis have been adapted for use in observational data, but risk-based methods for predictive HTE in observational data are still lacking and have been highlighted as an important future research need [17].

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: we compared the effect of angiotensin converting enzyme (ACE)

inhibitors to beta blockers on 9 outcomes across three different US claims databases.

2 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 databases 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 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 $O_1, \ldots, O_n$".

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 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

2.2 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 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) and the reliability of data entry.

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 [18,19]. To generate the target cohort (i.e. the set of patients on which the prediction model will be developed), we pool the already defined treatment cohort and comparator cohort. To avoid deferentially fitting the prediction model to patients across treatment arms, thus introducing spurious interactions with treatment [20,21], we develop the patient-level prediction model in the propensity score-matched (1:1) subset of the sample. 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.

After model development, 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.

2.4 Step 4: Estimation

We estimate 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 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 [22].

Prior to analyzing the results, we need to assess if adequate covariate balance was achieved within each risk stratum accounting for measured confounding. Common approaches include analysis of the overlap of propensity score distributions and calculation of standardized mean differences of the covariates before and after propensity score adjustment.

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

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.

2.6 Case study

As a demonstration of our framework, 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 3 efficacy outcomes (acute MI, hospitalization with heart failure and ischemic or hemorrhagic stroke) and 6 safety outcomes (hypokalemia, hyperkalemia, hypotension, angioedema, cough and abnormal weight gain). We used data from 3 US-based claims databases. The analysis plan was the framework outlined in steps 1 through 5.

A recent meta-analysis has shown that beta-blockers are on average less efficacious in preventing cardiovascular events than ACE inhibitors [23]. However, meta-analyses focusing on age of the patients have shown that beta blockers have similar efficacy in younger hypertensive patients, but should not be used as first-line treatment of hypertension in older patients mainly due to their increased risk of cardiovascular events compared to other first-line antihypertensive treatments [24,25].Our aim was to replicate these conclusions in our analyses.

3 Results

3.1 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 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 considered 3 efficacy and 6 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).

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 [4]. More information can be found in the supplementary material.

3.2 Step 2: Identification of the databases

For our analyses we used data from 3 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 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 MDCR respectively (Table 1). Adequate numbers of patients were included in all strata of predicted acute MI risk.

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 sample, using caliper of 0.2 and after excluding patients having the outcome any time prior to treatment initiation. We chose a time horizon of 2 years after inclusion into the target cohort. 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 (see supplementary material).

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 sample into 4 MI risk quarters. Within risk quarters, 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. The risk quarter-specific effect estimates were derived by averaging over the estimates within the propensity score fifths.

There was sufficient overlap of propensity score distribution in all risk strata (Figure 2). 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 [4]. Therefore, treatment effect estimates in the lowest risk quarters of CCAE and MDCR may still be subject to residual confounding and should be interpreted with care.

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 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 4). In MDCR hazard ratios remained stable from 0.93 (0.75 to 1.17; 95% CI) in the lowest MI risk quarter to 1.03 (0.92 to 1.16; 95% CI). 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 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.

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 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 -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).

3.6 Evaluation

Meta-analyses focusing on the age of the patients have found that beta blockers could only be considered as a first-line treatment for hypertension in younger patients, in whom risk of cardiovascular events is lower. The results of our case study are in line with this evidence, suggesting 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. Our case study however, was carried out as demonstration of the framework and more rigorous analyses are required to make any suggestions for clinical practice.

4 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 [16,17]. 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 3 efficacy and 6 safety outcomes. Our results were in line with the findings of relevant literature focusing on the age of patients treated for hypertension. 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

setting of RCTs [26]. 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 are usually not built with research in mind and therefore are susceptible to biases, poorly measured outcomes and missingness. All these can both obscure true treatment effect heterogeneity or falsely introduce it when there is none [7]. Therefore, inferences on both overall treatment effect estimates and treatment effect heterogeneity need to rely on strong, often unverifiable, assumptions, despite the advancements and guidance on best practices. However, it has been shown that well-designed observational studies on average replicate RCT results, even though often differences in magnitude may occur [31]. 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 [32].

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. Several sensitivity analyses have been proposed in the literature for measuring the robustness of results in the presence of unobserved confounding [33,34]. Another approach is to calibrate estimates and confidence intervals based on a large set of negative controls [35,36]. 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.

Ideally, externally derived and adequately validated prediction models would be preferred for analyzing treatment effect heterogeneity. 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 [20]. However, in observational databases treatment arms may significantly differ in sample size. Because the prediction model will possibly fit better 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. Other methods for achieving covariate balance before fitting the prediction model can be considered. For example, weighting on the propensity scores would be another valid approach.

Disease risk scores have been explored as an alternative to propensity scores for balancing covariates [37,38]. 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.

5 References

1 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.

2 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

3 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

4 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

5 Rothwell PM. Can overall results of clinical trials be applied to all patients? *The Lancet* 1995;**345**:1616–9. doi:10.1016/s0140-6736(95)90120-5

6 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

7 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

8 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

9 ROTHWELL P, MEHTA Z, HOWARD S, *et al.* From subgroups to individuals: General principles and the example of carotid endarterectomy. *The Lancet* 2005;**365**:256–65. doi:10.1016/s0140-6736(05)70156-2

10 Hayward RA, Kent DM, Vijan S, *et al.* 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

11 Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients. *JAMA* 2007;**298**:1209. doi:10.1001/jama.298.10.1209

12 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

13 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

14 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. doi:10.1161/circulationaha.105.558676

15 Sussman JB, Kent DM, Nelson JP, *et al.* Improving diabetes prevention with benefit based tailored treatment: Risk based reanalysis of diabetes prevention program. *BMJ* 2015;**350**:h454–4. doi:10.1136/bmj.h454

16 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

17 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

18 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

19 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. doi:10.7326/m14-0698

20 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

21 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

22 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

23 Wiysonge B CS, Opie L. Beta-blockers for hypertension. *Cochrane Database of Systematic Reviews* Published Online First: 2017. doi:10.1002/14651858.CD002003.pub5

24 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

25 Vögele A, Johansson T, Renom-Guiteras A, *et al.* Effectiveness and safety of beta blockers in the management of hypertension in older adults: A systematic review to help reduce inappropriate prescribing. *BMC Geriatrics* 2017;**17**. doi:10.1186/s12877-017-0575-4

26 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

27 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. doi:10.1056/nejm200006223422507

28 Ioannidis JPA. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286**:821. doi:10.1001/jama.286.7.821

29 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. doi:10.1093/eurheartj/ehs114

30 Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clinical Pharmacology & Therapeutics* 2017;**102**:924–33. doi:10.1002/cpt.857

31 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

32 Schuemie MJ, Ryan PB, Hripcsak G, *et al.* 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. doi:10.1098/rsta.2017.0356

33 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

34 Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2019;**82**:39–67. doi:10.1111/rssb.12348

35 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

36 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

37 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

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