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

1 Abstract (247 words)

Aim: One of the aims of the Observation 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 methods library with a framework for risk-based assessment of treatment effect heterogeneity.

Materials and Methods: 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.

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%.

Conclusions: 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.

2 Introduction

Interest in understanding how a treatment's effect varies across patients—a concept described as heterogeneity of treatment effects (HTE)—has been growing. This concept is central to the agenda for both personalized (or precision) medicine and comparative effectiveness research. More formally, HTE has been defined as 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]. However, even with well-established constant relative effects, treatment benefit (or harm) may vary substantially on the absolute scale.

More recently, "predictive" HTE analyses have been described (and contrasted with "one-variable-at-a-time" subgroup analysis) as approaches that provide predictions of potential outcomes in a particular patient with one intervention versus an alternative, taking into account multiple relevant patient characteristics. One promising approach is "risk modeling", in which treatment effects are estimated in strata of predicted risk [3,4]. Such a risk-based approach first stratifies patients according to baseline risk predictions, using either an existing or an internally developed risk prediction model [3]. Then, relative and absolute treatment effects are estimated within risk strata.

While these approaches have generally been recommended for application to clinical trials, 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 [5,6]. They are also often larger than many typical trials, providing excellent power for HTE analysis, and include heterogeneous populations. However, unlike trials, treatment effects are subject to confounding and the unique structure of different databases calls for database-specific analysis plans that are often not easily transportable.

The Observational Health Data Sciences and Informatics (OHDSI) collaborative (!!REF) 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) [7]. The common data structure enables analyses at a very large scale. For example, in a recent study [8], a large set of first-line treatments for hypertension was compared with respect to 55 outcomes in a network of databases, including 4.9 million patients from around the world.

We aimed to develop a framework for risk-based assessment of treatment effect heterogeneity in high-dimensional observational data. We implemented the framework using existing OHDSI methods for use in the OMOP-CDM, including the patient-level prediction framework [9] and the population-level effect estimation framework [10] based

on new-user cohort design. 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.

3 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 and evaluation of the results. A simple overview of the procedure can be seen in Figure XXXX.

3.1 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 $O_1, \ldots, O_n$". At least three cohorts are defined:

- 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 in their record.

3.2 Step 2: Identification of the database

The aim of this step is the inclusion of databases that represent the patient population of interest. The inclusion of multiple databases potentially increases the generalizability of results. Furthermore, the cohorts should preferably have adequate sample size to ensure precise effect estimation, even within smaller risk strata.

3.3 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 [11,12]. This prediction framework requires the definition of two essential cohorts: a target cohort and an outcome cohort.

To generate the target cohort we pool the already defined treatment cohort and comparator cohort . However, for risk-based analysis of treatment effects it is necessary to avoid deferentially fitting the prediction model to patients across the treatment arm in order to avoid inducing spurious interactions [13,14]. To do this, we developed the patient-level prediction model in the propensity score-matched subset of the population (1:1), where treatment assignment is well-balanced. The propensity scores are based on LASSO logistic regression for modeling the association between treatment assignment and all available demographics, drug exposures, diagnoses, measurements and medical procedures. Finally, we need to define the time horizon for which we aim to make predictions and we 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).

3.4 Step 4: Estimation

We use the patient-level prediction model to divide the target population into a set of equally-sized risk strata, typically 4 risk quarters. Then, we estimate propensity scores within risk strata. These propensity scores are used when estimating treatment effects, either by matching of patients from different treatment cohorts, by stratification of patients into groups with similar propensity scores, or by weighing patients' contribution to the estimation process. Within risk strata we estimate treatment effect both on the relative and the absolute scale. It is important to evaluate treatment effects in both scales, as effect cannot remain constant on both the relative and the absolute scale at the same time, assuming a non-zero treatment effect—treatment. 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.

3.5 Step 5: Result presentation and evaluation

Our framework provides standardized output for each step of the analysis. The number of patients and person years by treatment arm along with the number of outcomes. 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. This is rather relevant as 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 [15] demonstrated that the event rate and the discriminative ability of the prediction model can predict very well the distribution of predicted risk. The lower the event rate and the higher the 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.

Propensity score distributions by treatment group and covariate balance plots for each risk stratum. Event rates, hazard ratios and absolute risk differences in risk strata for a selected outcome, both in tables and in graphs. Hazard ratios and absolute risk differences for all analyzed outcomes by risk stratum. Finally, shiny application can be generated to enable easy sharing of the results.

4 Results

As a proof of concept, we focus on the comparison of angiotensin converting enzyme (ACE) inhibitors to beta blockers. ACE inhibitors are among the most common treatment classes for hypertension, with a 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 (!!REF). As a result, newer US guidelines do not consider them for initial treatment for hypertension while in the EU guidelines combination with other antihypertensive treatments is recommended [16,17]. 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 candidate for the treatment [18,19].

4.1 Step 1: General definition of the problem

We demonstrate the framework with the following research aim: "to 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 start 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 start 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 myocardial infarction (MI); hospitalization with heart failure; ischemic or hemorrhagic

stroke (efficacy outcomes); hypokalemia; hyperkalemia; hypotension; angioedema; cough; abnormal weight gain (safety outcomes).

All cohort definitions can be found in the supplementary material.

4.2 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.

4.3 Step 3: Prediction

To obtain a target cohort for developing patient-level predictions we first merged the ACE-inhibitors cohort with the beta blockers cohort. We then matched patients in the ACE-inhibitor cohort to patients in the beta blockers cohort on the propensity score. We focused on the efficacy outcomes (acute MI, hospitalization with heart failure and hemorrhagic or ischemic stroke) for risk stratification of the patient population. In each database, for each main outcome, we developed a prediction model. 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.

4.4 Step 4: Estimation

We used patient-level predictions to stratify the patient population into 4 risk quarters. We estimated relative and absolute treatment effects for all 9 outcomes of interest. In order to estimate risk quarter-specific treatment effects we first estimated propensity scores within risk quarters. We then used the propensity scores to stratify patients into 5 strata.

4.5 Step 5: Result presentation and evaluation

In the main manuscript we present the results of the analysis in the CCAE database with stratification based on risk predictions of acute MI. Results of analyses in the other databases and with other risk stratifications are included in the supplementary material.

For each outcome and in each risk stratum there were adequate numbers of patients (Table XXXX). The discriminative ability of the prediction models was moderate in the matched development subset (c-index 0.76 for acute MI ; 0.79 for hospitalization with heart failure; 0.74 for stroke;), in the general population (c-index 0.74 for acute MI; 0.77 for hospitalization with heart failure; 0.73 for stroke), in the treatment cohort (c-index for acute MI it was 0.71, for hospitalization with heart failure was 0.76 and for stroke it was 0.72) and in the comparator cohort (c-index for acute MI it was 0.79 for hospitalization with heart failure was 0.79 and for stroke it was 0.75).

Relative treatment effects of ACE-inhibitors vs beta blockers increased (hazard ratios decreased) with increasing acute MI risk, resulting in more pronounced increases of absolute treatment effects (ARD) with increasing acute MI risk (Figure XXXX). Patients in the low risk quarter did not receive absolute treatment benefit (ARD -0.03%) while absolute risk was 0.54% lower (95% confidence interval 0.36%—0.71%) for patients in the high-risk quarter. In contrast, the absolute and relative effects of ACE-inhibitors on safety outcomes (e.g. cough and angioedema) are approximately constant or even slightly decreasing with increasing acute MI risk (Figure XXXX and XXXX). Similar results were observed in the other two databases (see supplementary material).

This example nicely illustrates heterogeneity of absolute treatment effects, i.e. differences in absolute benefits and harms of ACE-inhibitors vs beta blockers for patients with different baseline risk. The 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. However, treatment with beta blockers may be a viable option in lower risk patients, in whom the benefit-harm tradeoff is in favor of beta blockers. This is in accordance with earlier findings that beta blockers should be considered as first-line treatment for younger hypertensive patients [18,20]. More thorough evaluation of these results is required in future research.

The results of the analyses performed can be accessed and assessed through a publicly available web application

(https://data.ohdsi.org/AceBeta9Outcomes).

5 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 [21]. However, low power and restricted prior knowledge on the mechanisms of variation in treatment effect are often inherent in RCTs, which are often 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 assessment of treatment effect heterogeneity. Multiple outcomes can be evaluated in patient subgroups of similar baseline outcome risk, where different outcome risk stratification schemes can be considered. The standardized nature of the framework enables transportability to multiple databases, provided that they are mapped to the OMOP-CDM.

Recently, guidelines on the application of risk modeling approaches for the assessment of heterogeneity of treatment effect in RCT settings have been proposed [22,23]. Our framework aims to translate these guidelines to the observational setting while also providing a toolset for its implementation. 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. This specific approach gave accurate results in extensive simulation studies [24]. However, such approaches do not account for unobserved confounding [25]. 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 [26,27]. 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 rough guide for clinical interpretation, it is not appropriate to guide clinical practice, where decisions need to be made at the individual level [22]. 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.

Externally derived prediction models are preferred for analyzing treatment effect heterogeneity [3]. Such external models should be well transportable. In the absence of such prediction models, simulations of RCTs have shown that internally derived models can be used to provide unbiased estimates of treatment effect across the spectrum of baseline risk [13]. 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, this may introduce spurious treatment-covariate interactions 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.

Our contribution is a translation of the PATH statement principles to the OHDSI methods library. Our methods encourage open science as it requires clear definition of the research questions translated into clear and reproducible cohort definitions that can easily be shared among researchers. Our R-package provides a standardized stepwise procedure for the assessment of HTE. This enables source code to be easily shared and evaluated. The results of these analyses can be reproduced in a straightforward manner. 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.

Recently, disease risk scores have been explored as an alternative to propensity scores for balancing covariates [28,29]. 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.

6 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 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

3 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

4 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

5 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

6 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

7 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

8 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

9 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

10 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

11 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

12 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

13 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

14 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

15 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

16 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

17 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

18 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

19 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

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

21 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

22 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

23 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

24 Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International Journal of Epidemiology* 2018;**47**:2005–14. doi:10.1093/ije/dyy120

25 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

26 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

27 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

28 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

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