Journal of the American Medical Informatics Association, 27(8), 2020, 1268–1277
doi: 10.1093/jamia/ocaa124
Research and Applications

Research and Applications

Large-scale evidence generation and evaluation across a network of databases (LEGEND): assessing validity using hypertension as a case study

Martijn J Schuemie [1,2] Patrick B Ryan,[1,3] Nicole Pratt,[4] RuiJun Chen [3,5] Seng Chan You,[6] Harlan M Krumholz,[7] David Madigan,[8] George Hripcsak,[3,9] and Marc A Suchard[2,10]

[1]Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA, [2]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, USA, [3]Department of Biomedical Informatics, Columbia University Medical Center, New York, New York, USA, [4]Quality Use of Medicines and Pharmacy Research Centre, University of South Australia, Adelaide, Australia, [5]Department of Medicine, Weill Cornell Medical College, New York, New York, USA, [6]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea, [7]Department of Medicine, Yale University School of Medicine, New Haven, Connecticut, USA, [8]Department of Statistics, Columbia University, New York, New York, USA, [9]Medical Informatics Services, New York-Presbyterian Hospital, New York, New York, USA, and [10]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, California, USA

Corresponding Author: Martijn J. Schuemie, PhD, Janssen Research and Development, 1125 Trenton Harbourton Rd, Titusville, NJ 08560, USA (schuemie@ohdsi.org)

Received 19 November 2019; Revised 2 April 2020; Editorial Decision 26 May 2020; Accepted 1 June 2020

ABSTRACT

Objectives: To demonstrate the application of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) principles described in our companion article to hypertension treatments and assess internal and external validity of the generated evidence.

Materials and Methods: LEGEND defines a process for high-quality observational research based on 10 guiding principles. We demonstrate how this process, here implemented through large-scale propensity score modeling, negative and positive control questions, empirical calibration, and full transparency, can be applied to compare antihypertensive drug therapies. We assess internal validity through covariate balance, confidence-interval coverage, between-database heterogeneity, and transitivity of results. We assess external validity through comparison to direct meta-analyses of randomized controlled trials (RCTs).

Results: From 21.6 million unique antihypertensive new users, we generate 6 076 775 effect size estimates for 699 872 research questions on 12 946 treatment comparisons. Through propensity score matching, we achieve balance on all baseline patient characteristics for 75% of estimates, observe 95.7% coverage in our effect-estimate 95% confidence intervals, find high between-database consistency, and achieve transitivity in 84.8% of triplet hypotheses. Compared with meta-analyses of RCTs, our results are consistent with 28 of 30 comparisons while providing narrower confidence intervals.

Conclusion: We find that these LEGEND results show high internal validity and are congruent with meta-analyses of RCTs. For these reasons we believe that evidence generated by LEGEND is of high quality and can inform medical decision-making where evidence is currently lacking. Subsequent publications will explore the clinical interpretations of this evidence.

© The Author(s) 2020. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

Key words: hypertension, treatment effects, observational studies, open science, empirical calibration

INTRODUCTION

The Observational Health Data Sciences and Informatics (OHDSI) Large-Scale Evidence Generation and Evaluation across a Network of Databases (LEGEND)[1–3] strives to produce reproducible evidence based on existing observational healthcare data, such as electronic health records (EHRs) and administrative claims data, and thus aims to fill in evidence gaps in medicine. Many studies document the limitations of much of the current observational research.[4,5] There are challenges related to selective reporting, nonreproducibility, confounding, imprecision, and lack of robust validation. As such, these studies are relegated to lower levels of evidence, and confidence in their ability to make causal inference is low. Meanwhile, evidence gaps in medicine are profound and the vast majority of recommendations in guidelines, even in the most evidence-based fields such as cardiology, are not supported by randomized trials. Clearly, we need more randomized trials, but, in parallel, there is a great need for credible observational evidence to support clinical decision-making.

The LEGEND design and execution is based on its 10 guiding principles,[2] aimed at addressing the current concerns about observational research. In brief, these principles prescribe the generation and dissemination of evidence on many research questions at once, for example, comparing all treatments for a disease for a wide range of outcomes, thus increasing the comprehensiveness of evidence and preventing publication bias.[6] These questions should be answered using a prespecified and systematic approach, preventing p-hacking (selective reporting). Best-practice statistical methods address measured confounding,[7] and control questions (research questions where the answer is known) quantify potential residual bias, expressed in calibrated confidence intervals (CIs)[8] and *P* values.[9] Finally, the evidence is generated in a network of databases to assess consistency, by sharing open source analytics code to enhance transparency and reproducibility but without sharing patient-level information, and ensuring patient confidentiality.

To demonstrate and evaluate LEGEND, in this article we apply the LEGEND principles to treatments for hypertension. Antihypertensive therapies carry well-established benefits in reducing blood pressure and the risk of major cardiovascular events. There remain large gaps in evidence about antihypertensive therapy—the health benefits and drug safety concerns of any 1 antihypertensive drug relative to other drugs as first-line therapy remain debatable—but among clinical areas, hypertension is less sparse than others. Reboussin et al[10] perform a systematic review of the current evidence from RCTs, and this review constitutes the basis of the recent 2017 American College of Cardiology / American Heart Association Guidelines[11] and the 2018 European Society of Cardiology and European Society of Hypertension Guidelines for the management of arterial hypertension.[12] However, the study by Reboussin et al presents data on only 40 head-to-head treatment comparisons and is based largely on studies completed before 2000.

We use the LEGEND methods to compare antihypertensive drugs and drug classes on effectiveness and safety outcomes. We start by illustrating our evidence generation process for a single research question. We then assess internal validity of the generated results for all the hundreds of thousands of research questions, based on LEGEND diagnostics defined in the Materials and Methods, and we use the meta-analysis of the Reboussin review as a benchmark to assess external validity. Discussion of the evidence on hypertension treatment itself, and the implications of this evidence for medicine, are beyond the scope of this paper and are instead presented in hypothesis-specific papers.[13]

MATERIALS AND METHODS

Figure 1 summarizes our approach based on the principles of LEGEND.[2] We define a large set of research questions and additionally define a set of control questions where the answer is known. We use effect estimates for the control questions to estimate systematic error distributions (eg, due to confounding, measurement error, and selection bias) and subsequent empirical calibration. We apply a systematic causal effect estimation procedure reflecting current best practices to generate estimates for all questions in an international network of healthcare databases. Each site runs the analysis locally and only shares aggregated statistics. The full result set is made available in an online database, accessible through various web applications. The protocol has been prespecified and made available online, alongside the open source code for executing the entire study.

Define a large set of research questions

Treatment comparisons

We analyze all pharmaceutical therapies indicated for hypertension treatment, as listed in the 2017 American Heart Association Guidelines,[11] categorized at 3 levels: drug ingredient, class, and major class (Table 1). Because medications are often prescribed in combinations, we also include all possible combinations of 2 treatments (either coprescribed individual drugs or combination products containing both ingredients). Similar to the guidelines, we focus on first-line therapies; we only evaluate the first hypertension treatment a patient receives and disregard all subsequent treatments. Only observed (at least 2500 new users in at least 1 database) treatments are considered and enumerated in Figure 1. For example, even though Table 1 lists 58 ingredients, we observe only 40 in the data and include those in the analysis. Similarly, even though we could hypothetically study $58 * (58 - 1) = 3306$ duo-ingredient therapies, only 66 are observed. We define our set of treatment comparisons as all possible (ordered) pairs of treatments.

Outcomes

Table 2 lists the 55 outcomes we include in our study. The primary effectiveness outcomes are acute myocardial infarction (AMI), hospitalization with heart failure, ischemic or hemorrhagic stroke, and a composite cardiovascular event outcome including the first 3 outcomes and sudden cardiac death. Additionally, we define a large set of safety outcomes based on known and suspected side effects of hypertension treatments that are listed on the product labels. The large number of treatment comparisons and outcomes not only allows us to answer more questions, but it also allows us to assess the operating characteristics of our design, answering questions such as whether we have an unexpected number of statistically significant results or whether there is an unusual pattern to the significance (eg, publication bias cuts at $P = .05$).[6]

We detect each outcome in the healthcare databases using carefully designed logic combining observed diagnosis, procedure, and

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

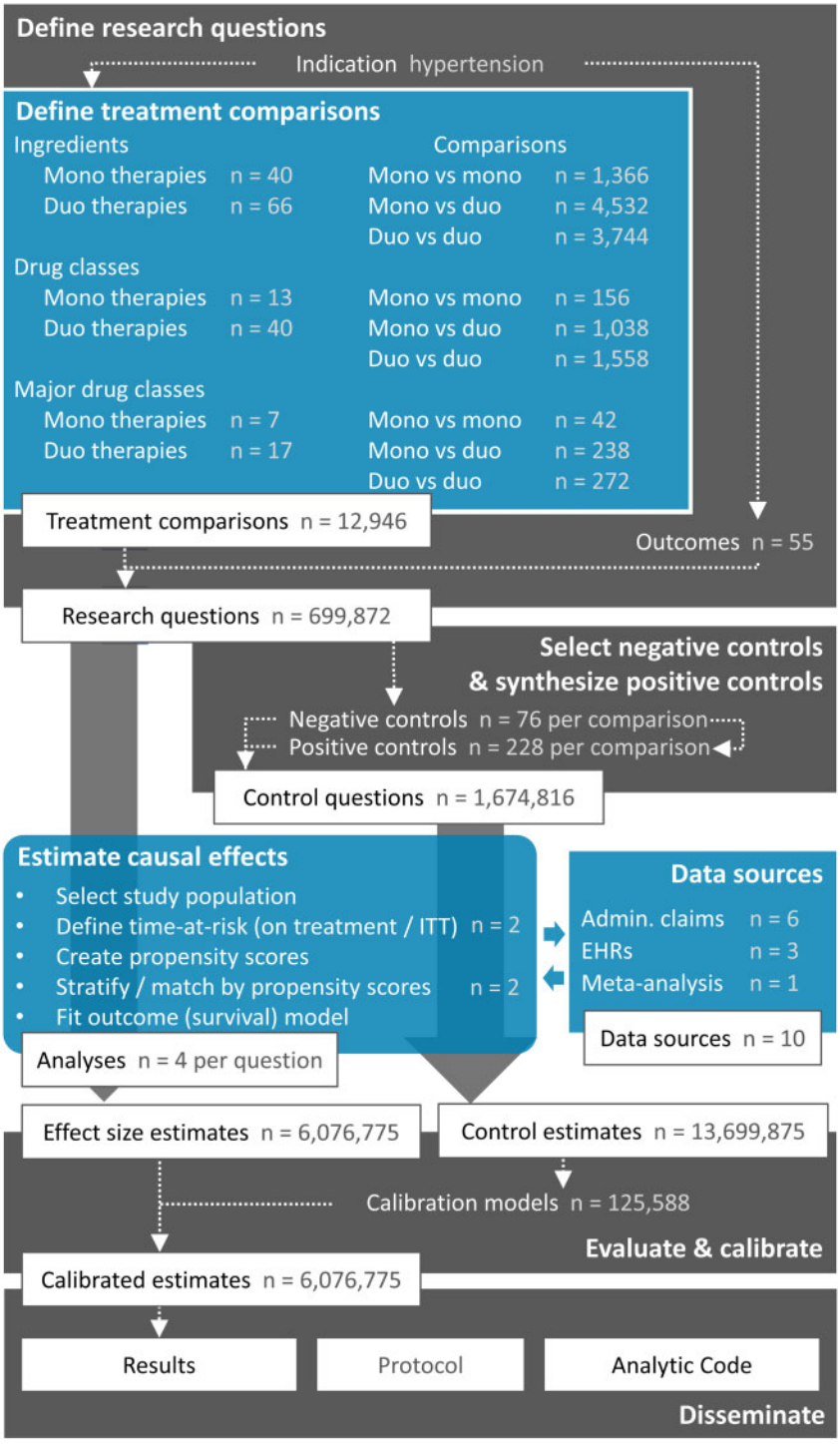

Figure 1. Overview of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) Hypertension study. The numbers reported here only include those data elements that were actually observed in the data. For exposures, a minimum of 2500 new users in at least 1 database was required. Outcomes had to be observed at least once. Admin. claims = administrative claims, EHRs = electronic health records, ITT = intent-to-treat.

other codes (see the protocol in the Supplementary Materials). For example, AMI is detected as the occurrence of an AMI diagnosis code in an emergency room or inpatient setting, with no AMI diagnosis in a similar setting in the prior 180 days. The AMI diagnosis codes are specified using OHDSI's Observational Medical Outcomes Partnership vocabulary standard concepts that map to the coding systems used in each of the databases (eg, 410.* in ICD-9).

Generate the evidence using best practices

For causal effect estimation, we compare a target treatment (T) to a comparator treatment (C) for the risk of an outcome (O). We define T and C as the first exposure to the treatments defined in Table 1, while requiring at least 1 year of prior observation, no prior hypertension treatment, and no other hypertension treatment starting within 7 days of treatment initiation. We further require a diagnosis

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

Table 1. Hypertension treatments included in this study

Ingredient		Class	Major class
Benazepril	Moexipril	ACE inhibitors	Angiotensin converting enzyme (ACE) inhibitors
Captopril	Perindopril		
Enalapril	Quinapril		
Fosinopril	Ramipril		
Lisinopril	Trandolapril		
Doxazosin	Terazosin	Alpha-1 blockers	Alpha-1 blockers
Prazosin			
Azilsartan	Losartan	Angiotensin receptor blockers	Angiotensin receptor blockers
Candesartan	Olmesartan		
Eprosartan	Telmisartan		
Irbesartan	Valsartan		
Atenolol	Bisoprolol	BB cardioselective	Beta-blockers (BB)
Betaxolol	Metoprolol		
Nebivolol		BB cardioselective and vasodilatory	
Carvedilol	Labetalol	BB combined alpha and beta receptor	
Acebutolol		BB intrinsic sympathomimetic activity	
Penbutolol	Pindolol		
Nadolol	Propranolol	BB non-cardioselective	
Amlodipine	Nicardipine	Dihydropyridine CCB (dCCB)	Calcium Channel Blockers (CCB)
Felodipine	Nifedipine		
Isradipine	Nisoldipine		
Diltiazem	Verapamil	Nondihydropyridine CCB (ndCCB)	
Hydralazine	Minoxidil	Direct vasodilators	Direct vasodilators
Eplerenone	Spironolactone	Aldosterone antagonist diuretics	Diuretics
Bumetanide	Torsemide	Loop diuretics	
Furosemide			
Amiloride	Triamterene	Potassium sparing diuretics	
Chlorthalidone	Indapamide	Thiazide or thiazide-like diuretics (THZ)	
Hydrochlorothiazide	Metolazone		
Aliskiren	Guanfacine		
Clonidine	Methyldopa		

Abbreviations: ACE, angiotensin converting enzyme; BB, beta blocker; CCB, Calcium Channel Blockers; dCCB, Dihydropyridine CCB; ndCCB, Nondihydropyridine CCB; THZ, thiazide.

Table 2. Health outcomes of interest

Abdominal pain	Dementia	Ischemic stroke
Abnormal weight gain	Depression	Malignant neoplasm
Abnormal weight loss	Diarrhea	Measured renal dysfunction
Acute myocardial infarction	End stage renal disease	Nausea
Acute pancreatitis	Fall	Neutropenia or agranulocytosis
Acute renal failure	Gastrointestinal bleeding	Rash
All-cause mortality	Gout	Rhabdomyolysis
Anaphylactoid reaction	Headache	Stroke
Anemia	Heart failure	Sudden cardiac death
Angioedema	Hemorrhagic stroke	Syncope
Anxiety	Hepatic failure	Thrombocytopenia
Bradycardia	Hospitalization with heart failure	Transient ischemic attack
Cardiac arrhythmia	Hospitalization with preinfarction syndrome	Type 2 diabetes mellitus
Cardiovascular event	Hyperkalemia	Vasculitis
Cardiovascular-related mortality	Hypokalemia	Venous thromboembolic events
Chest pain or angina	Hypomagnesemia	Vertigo
Chronic kidney disease	Hyponatremia	Vomiting
Cough	Hypotension	
Decreased libido	Impotence	

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

of hypertension recorded on or before the day of treatment initiation and no prior outcome O. We use 2 time-at-risk definitions: The on-treatment definition considers risk to start on the day after treatment initiation and end at treatment end, allowing for a maximum gap of 30 days between prescriptions. The intent-to-treat definition starts on the day after treatment initiation and stops at the end of observation. For each comparison, the study is restricted to the calendar time period when both treatments are observed in the database so that we compare drugs only during times when both are on the market.

We conduct our cohort study using the open-source OHDSI CohortMethod R package,[14] whose large-scale analytics are achieved through the Cyclops R package.[15] To account for the fact that treatment assignment is not random, resulting in imbalance between the target and comparator cohorts, we employ large-scale regularized regression to fit propensity models[16] using tens of thousands of baseline covariates.[7] These covariates include demographics, all prior drugs, conditions, procedures, etc. Hazard ratios are computed using Cox proportional hazards models conditioned on the propensity score-matched or stratified sets and are combined using meta-analysis for random effects. In addition, diagnostics on patient characteristic balance (ie, is every covariate in the propensity model balanced between the cohorts) and empirical clinical equipoise between exposure cohorts (is there sufficient overlap between the cohorts that an adjustment is possible, quantified using a preference score derived from the propensity score[16,17]) are generated.

Empirically evaluate through the use of control research questions

To diagnose and correct for residual confounding, we use control questions. Control questions are questions where the answer is known. We distinguish between negative controls, where the true hazard ratio is assumed to be 1, and positive controls with a known effect size greater than 1. We identify 76 negative controls—outcomes that are not believed to be caused by any hypertension treatment—through a data-rich algorithm[18] based on the literature, drug product labels, spontaneous reports, drug knowledge bases, and manual review (see the protocol in the Supplementary Materials). We use these to additionally generate $3*76=228$ synthetic positive controls with effect sizes 1.5, 2, and 4 by inserting simulated events into real data based on a prognostic model.[8] We estimate effect sizes for these control outcomes in each treatment comparison using the same evidence generation process as used for the research questions of interest. This allows us to evaluate the operating characteristics of our process (eg, how often the 95% CI contains the true effect size), and these characteristics are used to subsequently calibrate our CIs[8] and *P* values.[9]

Generate the evidence for all questions across a network of databases

We executed the LEGEND Hypertension study across the OHDSI research network[1] and included 9 databases covering 4 countries. Six databases contain administrative claims: IBM MarketScan Commercial Claims and Encounters, IBM MarketScan Medicare Supplemental Beneficiaries, IBM MarketScan Multi-state Medicaid, Optum ClinFormatics, Japan Medical Data Center, and the Korea National Health Insurance Service National Sample Cohort. Three databases contain electronic health records (EHRs): Optum deidentified Electronic Health Record Dataset (PanTher), Columbia University Medical Center, and QuintilesIMS Disease Analyzer Germany. In addition to the per-database effect size estimates, we also report summary estimates across databases using meta-analysis for random effects.[19] We compute the I^2 heterogeneity metric to assess between-database consistency.[20] An I^2 of zero means no between-database heterogeneity is observed.

Disseminate the generated evidence

We support open dissemination of generated evidence. Access to the database server containing the full results is available from the authors upon request. We have developed 2 web applications that connect to the database for exploring the results: The *LEGEND Basic Viewer* and *LEGENDMed Central*. The protocol and analytic code are posted publicly (https://github.com/OHDSI/Legend).

Internal validity

To assess internal validity, we examine the following properties.

Balance of the covariates demonstrates whether propensity score adjustment successfully creates comparison groups with similar characteristics. Our very strict goal is to achieve a standardized difference of the mean of 0.1 or less for every 1 of the thousands of measured covariates. Negative and positive controls allow us to assess residual confounding and *calibrate* CIs so that calibrated 95% CIs actually contains the true effect size 95% of the time. We report measured coverage of the calibrated CIs,[21] aiming for 95%.

The quantity I^2 is often used to test whether there is consistency among studies in a meta-analysis.[20] It is the percentage of total variation across studies that is due to heterogeneity rather than chance. We report *between-database consistency* as the proportion of hypotheses, after calibration, for which I^2 is below 0.25. We also graph the studies so that readers can inspect the overlap among estimates and among CIs.

We report *transitivity* as the proportion of studies for which superiority of 1 drug over a second, and that second drug over a third is accompanied by the finding that the first is superior to the third.

We assess whether balancing on a large number of covariates can improve balance on an *unmeasured confounder* using baseline blood pressure as an example. Blood pressure is not captured adequately in most of the databases included in LEGEND, except for the PanTher database. Using this database, we report covariate balance and the difference in effect size estimates with and without baseline blood pressure in the propensity model.

External validity

We compare the results of a meta-analysis across our databases to the direct meta-analysis of all RCTs included in the recent systematic review.[10] For each comparison, we compute the *P* value against the null hypothesis of no difference using:

$$z = \frac{log\ (HR_{RCT})\ - log(HR_{LEGEND})}{\sqrt{SE_{RCT}{}^2 + SE_{LEGEND}{}^2}}$$

$$p = 2{*}\Phi(-|z|)$$

Where HR_{RCT} and HR_{LEGEND} denote the hazard ratios of the RCT and LEGEND meta-analyses, respectively. SE_{RCT} and SE_{LEGEND} denote the standard errors of the RCT and LEGEND meta-analyses, respectively, and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. We report concordance between our results and those of the RCT meta-analysis as the number of comparisons where $P \geq .05$. We also inspect the statistical significance and direction of the results.

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

Table 3. Counts and maximum standardized difference per database. T = lisinopril, C = amlodipine, Outcome = angioedema, Max std. diff. = Maximum absolute standardized difference of means between T and C

	Subjects		Patient years		Outcomes			Max std. diff.	
Source	T	C	T	C	T	C	Covariate count	Before	After
CCAE	647 212	183 776	465 516	126 173	671	97	9765	0.407	0.029
MDCD	59 897	26 525	29 464	11 187	150	26	12 703	0.526	0.048
MDCR	73 821	32 375	61 864	28 107	99	18	11 217	0.363	0.029
Optum	447 905	143 079	340 148	107 631	475	88	11 903	0.432	0.028
PanTher	651 707	201 527	243 121	71 438	236	44	10 548	0.425	0.022
Total	**1 884 874**	**590 945**	**1 146 421**	**348 158**	**1646**	**279**			

Abbreviations: CCAE, Commercial Claims and Encounters; MDCD, Multi-state Medicaid; MDCR, Medicare Supplemental Beneficiaries; Optum, Optum ClinFormatics; PanTher, Optum deidentified Electronic Health Record Dataset.

RESULTS

In total, we analyze data from 21.6 million unique antihypertensive new users (the union of all unique patients entered in at least 1 Cox regression) to generate 6 076 775 effect size estimates answering 699 872 research questions, each including full diagnostics and all additional validity checks.

Exemplar research question: The effect of lisinopril compared to amlodipine on the risk of angioedema

For illustration, we highlight just 1 of the 699 872 research questions: the effect of lisinopril compared to amlodipine on the risk of angioedema. Table 3 reports the numbers of patients initiating treatment of 1 of these drugs (monotherapy) in each database and includes the number of angioedema events observed during the on-treatment time-at-risk. We furthermore report the number of covariates constructed in each database for these patients, and the maximum absolute standardized difference of the mean across covariates before and after propensity score matching. As shown, all of the thousands of covariates had an absolute standardized difference of the mean smaller than 0.1, indicating good balance. Note that some databases do not have sufficient exposure to both drugs and are therefore omitted.

Figure 2 reports the estimated hazard ratios for the risk of angioedema with lisinopril compared to amlodipine, for each database separately as well as a summary estimate. This same procedure is used not only for angioedema, but also for the 304 control outcomes. Using the estimates for these control questions, we fit a systematic error model and use it to compute the empirically calibrated estimates reported in Figure 2.

The results above use the on-treatment time-at-risk window and propensity score matching, producing a calibrated hazard ratio of 3.11 (2.36–4.48). When using an intent-to-treat window instead, the calibrated summary hazard ratio is 1.92 (1.57–2.40). Using propensity score stratification, the calibrated summary hazard ratio is 2.52 (1.95–3.32) and 1.59 (1.31–1.95) for the on-treatment and intent-to-treat window, respectively.

Internal validity across all research questions

We now report the internal and external validity checks across all research questions examined in this LEGEND hypertension study.

Balance

When using propensity score matching, we achieve balance (standardized difference of the means < 0.1 on every 1 of the thousands of included covariates) for 75% of the 1 665 176 effect size estimates. When using propensity score stratification this is 19% of 2 280 73 estimates. Note that we can compute fewer estimates when using matching compared to stratification because matching may lead to removal of subjects for which no match could be found.

Empirical evaluation and calibration

Of the 13 699 875 control estimates, 79.9% contain the true effect size within the 95% CI before calibration. That is, according to our negative and positive controls, when we calculate a 95% CI, that interval includes the true value only 79.9% of the time, implying more than an expected number of false-positive results would be reported. After our empirical calibration the coverage becomes 95.7% (nominal = 95%).

Between-database consistency

We identify 37 953 target-comparator-outcome triplets having sufficient data in at least 4 databases to compute an estimate for the analysis specifying an on-treatment time-at-risk using propensity score matching. Across databases, 78% of calibrated estimates have an I^2 below 0.25, corresponding to low heterogeneity.[20] That is, in 78% of our eligible comparisons, the databases were consistent with each other under the most conservative threshold (ie, 0.25). In contrast, only 64% of the estimates have an I^2 below 0.25 when no calibration is applied. The I^2 score is computed and available for all meta-analyses in the results database. In this manuscript, we report all meta-analysis estimates irrespective of I^2.

Transitivity

If treatment A has a statistically significant higher risk than treatment B for a particular outcome, and treatment B has a statistically significant higher risk than C for that same outcome, we expect A to have a statistically significant higher risk than C. In total, we identify 653 595 such A-B-C combinations in the calibrated meta-analyses results, of which for 554 268 triplets (84.8%) the transitivity property holds.

Effect of not explicitly adjusting for baseline blood pressure

Supplementary Table S1 evaluates the balance on blood pressure before and after propensity score matching for a select number of treatment comparisons in the PanTher database. Before matching, there is an absolute standardized difference of means greater than 0.1 for almost all comparisons, suggesting the treatment groups have different baseline blood pressure. After matching on propensity scores using the original set of covariates—which excludes blood

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

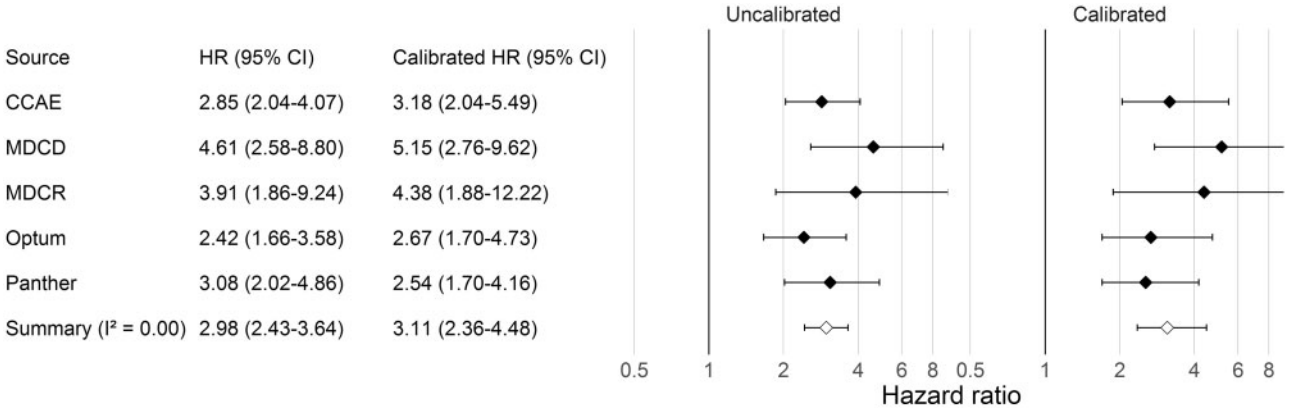

Figure 2. Hazard ratio (HR) estimates (and 95% CIs) before and after empirical calibration for lisinopril compared to amlodipine for the risk of angioedema, when using propensity score matching and an on-treatment time-at-risk window.

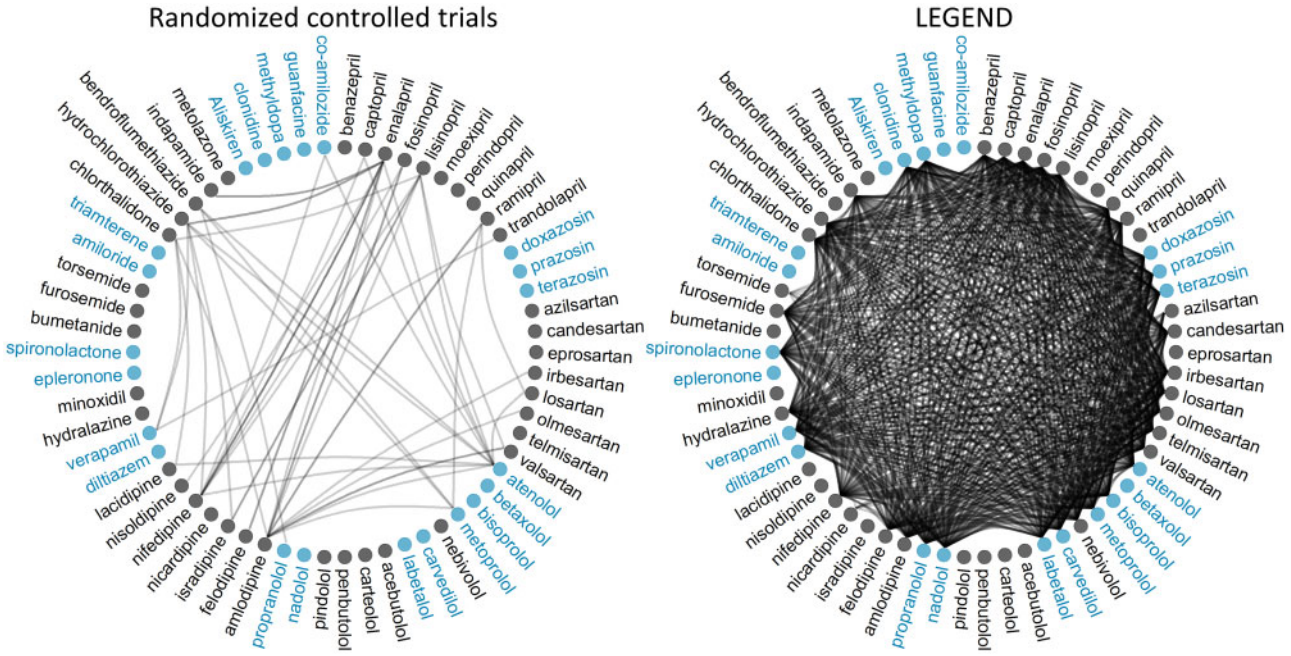

Figure 3. Comparisons of single-drug hypertension treatments in randomized controlled trials (left) and in LEGEND (right). Each circle represents an ingredient. Color groupings indicate drug classes. A line between circles indicates the 2 drugs are compared in at least 1 study.

pressure—we see a reduced imbalance in blood pressure, but a few comparisons cross our predefined threshold of 0.1. To evaluate whether these residual imbalances should cause bias, we repeat our analysis including blood pressure in the propensity model. Supplementary Table S1 shows that after matching on these propensity scores, baseline blood pressure is nearly perfectly balanced. Supplementary Figure S1 shows that use of the 2 different propensity scores has little to no impact on the hazard ratio estimates produced.

External validity

We evaluate external validity by comparing the LEGEND results to meta-analyses of randomized controlled trials (RCTs). All RCTs included in the recently published systematic review cover the 40 head-to-head comparisons of drugs shown in the left of Figure 3.[10] In contrast, LEGEND analyzes 12 946 treatment comparisons for each outcome. The mono-ingredient versus mono-ingredient comparisons performed are shown in the right of Figure 3. The sample size (subjects) for comparisons in published RCTs varied from 102 to 33 000,[10] with a median of 1148. In contrast, the sample size for comparisons in LEGEND (excluding meta-analyses) varied from 692 to 1.2 million, with a median of 33 771.

Concordance

Of the 40 comparisons reported in the systematic review, 30 overlapped with research questions addressed in LEGEND. Of these 30 comparisons, 28 (93%) show no statistical difference in estimates,

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

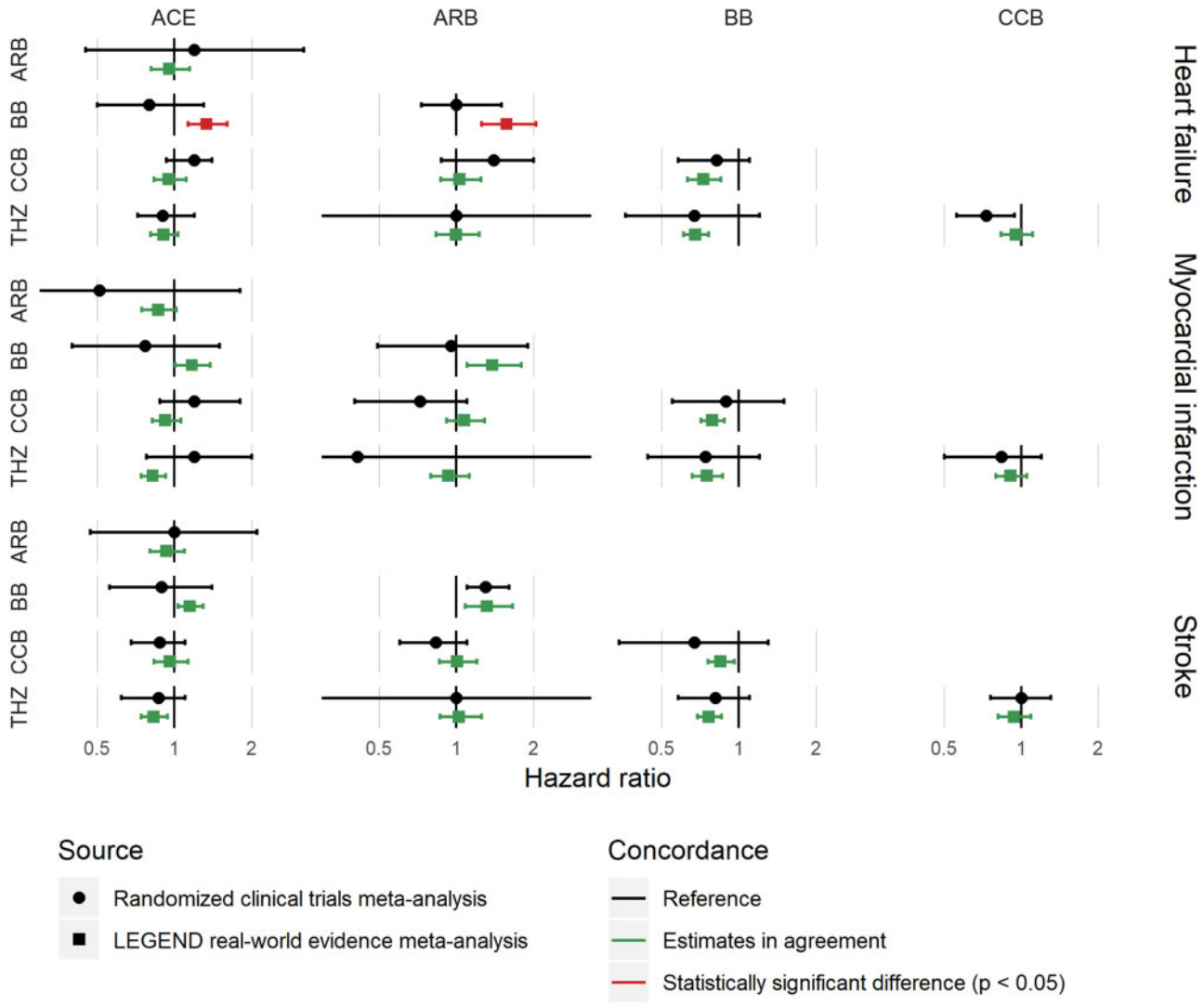

Figure 4. Concordance between LEGEND meta-analysis results (using propensity score matching with an on-treatment risk window as well as empirical calibration) and the results from meta-analyses of randomized controlled trials. ACE = ACE inhibitors, ARB = Angiotensin receptor blockers, BB = Beta-blockers, CCB = Calcium channel blockers, THZ = Thiazide or thiazide-like diuretics. Hazard ratios greater than 1 indicate greater risk for the drug class at the left.

and 2 (7%) have a nominal P value $< .05$ indicating different estimates. Note that these numbers do not correct for multiple testing, so 5% are expected to have $P < 0.05$ when no real differences exist. Figure 4 shows, for 3 effectiveness outcomes, the concordance between results from LEGEND and direct meta-analyses of RCTs.[10] Of the 30 comparisons, 16 agreed on the significance and direction of the result (both not significant or both significant in the same direction), 1 was significant in the systematic review but not significant in LEGEND, and 13 were significant in LEGEND but not in the systematic review. The interpretation of this is in the Discussion.

DISCUSSION

In this study we apply the LEGEND principles to systematically compare all pharmaceutical therapies indicated for hypertension treatment, considering a wide range of effectiveness and safety outcomes. We find that these LEGEND results have high internal validity and are statistically congruent with available meta-analyses of RCTs, although with its greater sample size, LEGEND often found statistical significance where the meta-analyses did not.

For internal validity, we find that propensity score matching achieves balance on every 1 of the thousands of covariates for 75% of effect size estimates, despite our very stringent criterion (absolute standardized difference < 0.1), which is normally used in studies with only a handful of covariates.[22] Even without empirical calibration, coverage of the 95% CIs remained relatively high at 79.9% compared to the 6%–88% observed in prior research.[8] After calibration, this coverage became almost identical to the nominal 95%. High consistency between databases, as expressed in low I^2 scores, suggests the results are robust and not due to database idiosyncrasies. This is even more remarkable when considering the heterogeneity of the populations across the OHDSI research network, which included EHR and administrative claims data from 4 countries. We observe a transitivity of 84.8%, but this number is hard to interpret. To our knowledge, no one has ever attempted to quantify transitivity in scientific results before. We would not expect a transitivity of 100% because of varying standard errors in our meta-analytic esti-

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

mates (which assume random effects), and because for each comparison we restrict to the calendar time when both drugs are on the market.

For external validity, we compare the LEGEND results to a set of direct meta-analyses of RCTs, which were used to generate recent hypertension treatment guidelines. LEGEND results showed high concordance with the meta-analysis (Figure 4), showing a statistically significant difference from the meta-analysis at about the rate one would expect based on random error alone, if no real differences exist (7% observed, 5% expected). Because of the larger sample size, LEGEND results tend to have narrower CIs. In some cases, the LEGEND point estimate can be on the opposite side of 1 from the RCTs point estimate, and still be concordant if their CIs overlap sufficiently. We emphasize that when a meta-analysis CI includes 1, the formal conclusion is not "no effect" but instead insufficient evidence to conclude an effect. Therefore, even if LEGEND shows a statistically significant effect where the RCT meta-analysis demonstrates no effect (eg, beta blockers compared to thiazides for all 3 outcomes), they may still be consistent. In fact, this is 1 of the goals of LEGEND, to use its larger sample size to more precisely measure effects and to uncover effects masked by the RCTs' small sample sizes. Therefore, while we do not know the underlying true effect sizes and cannot credit LEGEND for getting a better estimate than the meta-analysis, similarly we cannot count it against LEGEND that it had a narrower interval and apparently uncovered an effect. The best we can say is that the meta-analysis and LEGEND overlapped sufficiently (ie, sufficiently that they could have been drawn from the same theoretical population of studies). Of the 2 hypotheses that were statistically significant in the meta-analysis, LEGEND was statistically consistent with both (ie, the CIs were overlapping enough) but in 1 of the 2, LEGEND was not statistically significant. Here again, in theory, the meta-analysis and LEGEND could have been drawn from the same theoretical population of studies, just as RCTs for the same comparison sometimes differ in statistical significance. We recognize that merely guessing no effect with a wide CI would have captured all the RCT meta-analysis results, producing even better concordance, but LEGEND did not in fact do that and instead produced narrower intervals that overlapped the meta-analysis results.

The reason for the discordance between LEGEND and the meta-analysis on beta-blockers versus angiotensin-converting enzyme inhibitors and versus angiotensin receptor blockers remains unclear. It could be chance, residual bias in the observational study, or a difference in populations (RCTs tend to be run on sicker patients with a history of hypertension treatment, sometimes on several drugs, whereas this LEGEND study was focused on first-time use of an antihypertensive drug).

The LEGEND meta-analyses provide estimates for 699 872 unique research questions and can thus help inform clinical decision-making when other information is not available. Reviewing all these other results goes beyond what can be discussed in this article. We are in the process of writing several clinical papers on specific questions likely to be of high interest, and have already published a paper based on these LEGEND results comparing first-line hypertension treatments at the class level.[13] This article demonstrates that even though the current guidelines[11] do not distinguish between a wide set of drugs as the recommended choice for first-line treatment, there are differences in the effectiveness and safety of these drugs that warrant consideration. Clinical researchers and appropriately trained clinicians can consult the web apps themselves to seek answers for specific questions they have. Each LEGEND result comes with full diagnostics, including a description of the study population, key characteristics, propensity score distribution, covariate balance, bias distribution as estimated using negative and positive controls, and Kaplan–Meier plots. We invite others to develop other apps or to help interpret these results in other ways.

LEGEND answers each question using best practices, including advanced methods for adjusting for confounding, producing study diagnostics, and replication across a network of databases. For these reasons, we believe the evidence LEGEND generates, in this case for treatments of hypertension, is of high quality, and can inform medical decision-making where evidence is currently lacking.

LEGEND uses large-scale propensity scores as the primary means to adjust for confounding. The technique has shown promise[7] and, in this experiment, demonstrated that it adjusted for an important potential confounder, baseline blood pressure, without including it in the model. Other methods are also possible, such as use of a more traditional (ie, not large-scale) propensity score, but incorporating a prognostic score[23] to produce an approach that may be doubly robust. Comparing these techniques is out of scope for this article, but further research in this area is warranted.

Limitations

Like all observational research, LEGEND is still vulnerable to residual bias due to confounding and measurement error. However, in contrast to the vast majority of published observational studies, LEGEND provides a rich set of diagnostics to evaluate whether we can trust the results, including covariate balance and estimates for control questions. Any systematic error observed through the control questions is incorporated in calibrated CIs and *P* values, thus conveying the limits of what can be learned from these data.

One limitation of using existing healthcare data for research purposes is that some variables of interest may not be recorded systematically. In this study, the most prominent example is blood pressure, which ideally would have been included in the propensity score, but is not available in all databases. Our sensitivity analysis in PanTher indicates that access to these data does not meaningfully change the effect size estimates, suggesting that baseline blood pressure is either not an important confounder or, more likely, is already sufficiently adjusted for by the large-scale propensity scores through observed proxy variables. Similarly, postintervention blood pressure could have been included as an outcome, but, because of its unavailability in the databases, we did not use it.

We purposely do not correct for multiple hypotheses in our results because that can only be done once researchers choose a specific hypothesis or set of hypotheses. As when using results from the literature, it is important to consider false positives when faced with multiple testing, and our results readily allow for adjustment for multiple testing because we disseminate all results.

CONCLUSION

We find that the LEGEND results have high internal validity and are congruent with direct meta-analyses of RCTs. Even though many RCTs inform on the effects of hypertension treatments, much uncertainty remains. By following the LEGEND guiding principles that address study bias, p-hacking, and publication bias, LEGEND seeks to augment existing knowledge by generating reliable evidence from existing healthcare data, answering hundreds of thousands of research questions simultaneously using a transparent, reproducible, and systematic approach.

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020

FUNDING

This work was supported in part by National Science Foundation grant IIS 1251151, National Institutes of Health grant R01 LM006910, and Australian National Health and Medical Research Council grant GNT1157506.

AUTHOR CONTRIBUTIONS

Each author made substantial contributions to the conception or design of the work; was involved in drafting the work or revising it critically for important intellectual content; gave final approval of the version to be published; and has agreed to be accountable for all aspects of the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Schuemie MJ, Hripcsak G, Ryan PB, *et al*. Empirical CI calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci USA* 2018; 115 (11): 2571–7.
2. Schuemie MJ, Ryan PB, DuMouchel W, *et al*. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Stat Med* 2014; 33 (2): 209–18.
3. Hripcsak. G, Duke JD, Shah NH, *et al*. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
4. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986; 7 (3): 177–88.
5. Higgins JPT, Thompson SG, Deeks JJ, *et al*. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327 (7414): 557–60.
6. Schuemie MJ, Ryan PB, Pratt N, *et al*. Principles of Large-Scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *J Am Med Inform Association*. 27(8):1331–1337.
7. Schuemie MJ, S, Cepede M, Suchard MA, *et al*. *How Confident Are We about Observational Findings in Health Care: A Benchmark Study. 2.1*. 2020; doi: 10.1162/99608f92.147cc28e.
8. Schuemie MJ, Suchard MA, Ryan PB. CohortMethod: New-user cohort method with large scale propensity and outcome models. 2018. https://ohdsi.github.io/CohortMethod/
9. Williams B, Mancia G, Spiering W, *et al*. ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of *Hypertension. J Hypertens* 2018; 36 (10): 1953–2041.
10. 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. *Lancet*. 2019; 394 (10211): 1816–26.
11. Reboussin DM, Allen NB, Griswold ME, *et al*. Systematic review for the 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. *Circulation* 2018; 138: e595–616.
12. 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: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 2018; 71 (6): 1269–324.
13. Suchard MA, Simpson SE, Zorych I, *et al*. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Trans Model Comput Simul* 2013; 23 (1): 1–17.
14. Rosenbaum PR, Rubin DB. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*.Biometrika 1983;70:41–55.; doi: 10.21236/ada114514.
15. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol* 2018; 47 (6): 2005–14.
16. Walker A, Lauer Patrick, *et al*. A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effect Res* 2013; 3: 11–20.
17. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009; 28 (25): 3083–107.
18. Hansen BB. The prognostic analogue of the propensity score. *Biometrika* 2008; 95 (2): 481–8.

Downloaded from https://academic.oup.com/jamia/article/27/8/1268/5895564 by guest on 13 November 2020