*Journal of the American Medical Informatics Association*, 27(8), 2020, 1331–1337
doi: 10.1093/jamia/ocaa103
Perspective

Perspective

# Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND)

**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, 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, South Australia, 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, California, 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, [10]Department of Biomathematics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, California, USA

Corresponding author: Martijn J. Schuemie, PhD, 1125 Trenton Harbourton Rd, Titusville, NJ 08560, USA (schuemie@ohdsi.org)

Received 2 December 2019; Revised 27 March 2020; Editorial Decision 14 May 2020; Accepted 16 May 2020

**ABSTRACT**

Evidence derived from existing health-care data, such as administrative claims and electronic health records, can fill evidence gaps in medicine. However, many claim such data cannot be used to estimate causal treatment effects because of the potential for observational study bias; for example, due to residual confounding. Other concerns include *P* hacking and publication bias.

In response, the Observational Health Data Sciences and Informatics international collaborative launched the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) research initiative. Its mission is to generate evidence on the effects of medical interventions using observational health-care databases while addressing the aforementioned concerns by following a recently proposed paradigm. We define 10 principles of LEGEND that enshrine this new paradigm, prescribing the generation and dissemination of evidence on many research questions at once; for example, comparing all treatments for a disease for many outcomes, thus preventing publication bias. These questions are answered using a prespecified and systematic approach, avoiding *P* hacking. Best-practice statistical methods address measured confounding, and control questions (research questions where the answer is known) quantify potential residual bias. 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.

Here we detail the LEGEND principles and provide a generic overview of a LEGEND study. Our companion paper highlights an example study on the effects of hypertension treatments, and evaluates the internal and external validity of the evidence we generate.

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

## INTRODUCTION

Real-world evidence derived from existing health-care data, such as administrative claims and electronic health records (EHRs), can fill evidence gaps in medicine. However, such observational research for the purpose of causal inference is often criticized because of the potential for bias.[1, 2] The main reason cited is confounding: observational studies are prone to detect spurious effects because treatment is not assigned randomly, and 1 treatment group may therefore fundamentally differ from another in ways that affect the outcome risk. Even though observational studies often attempt to correct for this, not all confounders may be known, measured, or adjusted for correctly. Other concerns with observational research include the issues of *P* hacking and publication bias. *P* hacking occurs when a researcher performs multiple variations of the analysis until the desired result is obtained, while publication bias occurs when journals selectively publish "statistically significant" results or authors only choose to submit studies with positive effects. Both *P* hacking and publication bias can increase the false positive rate in published research, due to the hidden multiple testing.[3]

To address these concerns, informaticists, statisticians, and clinicians from the Observational Health Data Sciences and Informatics (OHDSI) international collaborative [4] launched the Large-Scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) research initiative. LEGEND generates evidence on the effects of medical interventions using observational health-care data, while addressing the aforementioned concerns of observational research by following a recently proposed paradigm for generating evidence.[5] A key element of this new paradigm is that evidence should be generated at scale. Rather than answering a single question at a time, many research questions are better addressed in a single study; for example, comparing all treatments for a disease for a wide range of outcomes of interest. This shift to large-scale analyses enhances the comprehensiveness of the evidence base, and disseminating all generated evidence without filtering prevents publication bias and *P* hacking. Furthermore, applying a systematic approach to answer these questions allows us to evaluate the performance of our evidence generation process. Traditionally, each observational study answers a single question using an ad hoc design with unknown operating characteristics, making it unclear to what extent the results of these studies can be trusted. In LEGEND we include control questions—questions where the answer is known—to measure operating characteristics, and use this information to calibrate our confidence intervals (CIs) and *P* values. Lastly, by performing this analysis in a network of heterogeneous, observational, health-care databases, we can observe whether findings in 1 database replicate in other databases, thus enhancing the reproducibility of the findings.

This paper describes the guiding principles of LEGEND (see Table 1) that enshrine this new paradigm, and provides an overview of a typical LEGEND study. A companion paper [6] describes an example study on the effects of hypertension treatments, and explores the internal and external validity of the evidence we generate. Future LEGEND studies will generate evidence for other disease areas.

## MATERIALS AND METHODS

### Guiding principles

Principles 1 and 2, together, prevent publication bias, and 3 and 4 address *P* hacking. Principles 5, 6, and 8 aim to minimize the impact of biases associated with observational studies, first by using advanced methods to correct for observed confounding, and second by using control questions to measure residual bias after these corrections and to calibrate statistics accordingly. Principle 7 enhances the transparency of results, Principle 9 addresses the generalizability of pooled results by examining the heterogeneity of effect estimates across databases, and Principle 10 addresses data security and privacy. By applying these principles, we therefore overcome the biggest concerns for observational research and enhance confidence in the application of observational research for clinical decision-making.

### Overview of a LEGEND study

Figure 1 shows an overview of a typical LEGEND study.

We start by defining a large set of research questions (Principle 1), as well as a set of control questions where the answer is known (Principle 6). We apply a systematic, causal effect estimation procedure reflecting current best practices (Principles 5 and 8) to generate estimates for all questions (Principle 4) in an international network of health-care databases (Principle 9). Each site runs the analysis locally and only shares aggregated statistics (Principle 10). We use effect estimates for the control questions to estimate systematic error distributions (for example, due to confounding, measurement error, and selection bias) and subsequent empirical calibration. The full result set is made available in an online database, accessible through various web applications (Principle 2). The protocol has been prespecified and made available online (Principle 3), alongside the open-source code for executing the study (Principle 7).

### Define a large set of research questions (Principle 1)

We predefine the set of treatments we wish to compare; for example, all treatments for a particular indication (eg, all treatments for hypertension). We define the set of treatment comparisons as all (ordered) pairs of treatments, and specify the set of health outcomes of interest, which may include both efficacy and safety outcomes. Our set of research questions is then defined by the combination of each treatment pair with each outcome of interest.

For example, if for some indication we identify 10 different treatments, we can construct $10 * (10 - 1) = 90$ treatment pairs. If we further specify 20 outcomes of interest, we can define $90 * 20 = 1800$ research questions. Importantly, this set of research questions, as well as the full study design, are specified before the analysis is executed and are posted publicly, as described in the section on Transparency.

### Empirically evaluate through the use of control research questions (Principle 6)

To determine the potential for systematic bias in each treatment comparison, a series of control questions is defined. Control questions are questions where the answer is known, and can be either negative controls, where the true hazard ratio is assumed to be 1, or positive controls, with a known effect size greater than 1.

Negative controls or "falsification hypotheses" have been proposed as a diagnostics tool for observational studies.[9–11] When comparing 2 treatments, we specify negative controls as selected outcomes that are not believed to be caused or prevented by either treatment. For example, neither amlodipine nor lisinopril are believed to cause or prevent ingrown nails. When comparing these 2 hypertension treatments, we therefore assume that the hazard ratio for ingrown nails is equal to 1. Even though there is no causal relationship from either drug to ingrown nails, the relationship may very well be confounded; for example, because ingrown nails tend to occur more often in the elderly and 1 of the treatments is also

**Table 1**: Guiding principles of the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) initiative.

| | |
|---|---|
| 1 | **LEGEND will generate evidence at a large scale.**<br>Instead of answering a single question at a time (eg, the effect of 1 treatment on 1 outcome), LEGEND answers large sets of related questions at once (eg, the effects of many treatments for a disease on many outcomes).<br>**Aim:** Avoids publication bias, achieves comprehensiveness of results, and allows for an evaluation of the overall coherence and consistency of the generated evidence. |
| 2 | **Dissemination of the evidence will not depend on the estimated effects.**<br>All generated evidence is disseminated at once.<br>**Aim:** Avoids publication bias and enhances transparency. |
| 3 | **LEGEND will generate evidence using a prespecified analysis design.**<br>All analyses, including the research questions that will be answered, will be decided prior to analysis execution.<br>**Aim:** Avoids *P* hacking. |
| 4 | **LEGEND will generate evidence by consistently applying a systematic process across all research questions.**<br>This principle precludes modification of analyses to obtain a desired answer to any specific question. This does not imply a simple one-size-fits-all process, rather that the logic for modifying an analysis for specific research questions should be explicated and applied systematically.<br>**Aim:** Avoids *P* hacking and allows for the evaluation of the operating characteristics of this process (Principle 6). |
| 5 | **LEGEND will generate evidence using best practices.**<br>LEGEND answers each question using current best practices, including advanced methods to address confounding, such as propensity scores. Specifically, we will not employ suboptimal methods (in terms of bias) to achieve better computational efficiency.<br>**Aim:** Minimizes bias. |
| 6 | **LEGEND will include empirical evaluation through the use of control questions.**<br>Every LEGEND study includes control questions. Control questions are questions where the answer is known. These allow for measuring the operating characteristics of our systematic process, including residual bias. We subsequently account for this observed residual bias in our *P* values, effect estimates, and confidence intervals using empirical calibration.[7,8]<br>**Aim:** Enhances transparency on the uncertainty due to residual bias. |
| 7 | **LEGEND will generate evidence using open-source software that is freely available to all.**<br>The analysis software is open to review and evaluation, and is available for replicating analyses down to the smallest detail.<br>**Aim:** Enhances transparency and allows replication. |
| 8 | **LEGEND will not be used to evaluate new methods.**<br>Even though the same infrastructure used in LEGEND may also be used to evaluate new causal inference methods, generating clinical evidence should not be performed at the same time as method evaluation. This is a corollary of Principle 5, since a new method that still requires evaluation cannot already be best practice. Also, generating evidence with unproven methods can hamper the interpretability of the clinical results. Note that LEGEND does evaluate how well the methods it uses perform in the specific context of the questions and data used in a LEGEND study (Principle 6).<br>**Aim:** Avoids bias and improves interpretability. |
| 9 | **LEGEND will generate evidence across a network of multiple databases.**<br>Multiple heterogeneous databases (different data capture processes, health-care systems, and populations) will be used to generate the evidence to allow an assessment of the replicability of findings across sites.<br>**Aim:** Enhances generalizability and uncovers potential between-site heterogeneity. |
| 10 | **LEGEND will maintain data confidentiality; patient-level data will not be shared between sites in the network.**<br>Not sharing data will ensure patient privacy, and comply with local data governance rules.<br>**Aim:** Privacy. |

*Note:* LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.

preferentially prescribed in the elderly. We use the negative controls to evaluate whether we indeed produce an estimate of no effect after we implement our strategies for confounding control. To select negative controls, information from literature, product labels, and spontaneous reporting can automatically be extracted and synthesized to produce a candidate list of outcomes with no known links with the treatments of interest.[12] The candidate list can be rank ordered by prevalence of the outcome and manually reviewed to determine whether they are appropriate to be included. For every research question, it may be necessary to identify new sets of negative controls, although sometimes these may be reused where appropriate.

Positive controls are also employed to detect some types of bias not captured by negative controls alone, such as bias towards the null. Because unlike negative controls, we seldom know the true effect size of real positive controls, we employ synthetic positive controls, constructed by adding simulated outcomes to real negative controls.[8] These simulated outcomes are only inserted during exposure to 1 of the treatments, thus artificially increasing the effect size. To preserve observed confounding, the new outcomes are sampled from predicted probabilities based on baseline patient characteristics. Note that some types of bias, such as bias due to unmeasured confounding and differential misclassification of outcomes caused by the exposure, will not be captured by these positive controls. From each negative control (with the true effect size of 1), positive controls should be generated to simulate various effect sizes (eg, effect sizes 1.5, 2, and 4).

Through our control questions, we evaluate whether our process produces results in line with known effect sizes. Importantly, we estimate the CI coverage probability: the proportion of time that the CI contains the true value of interest. For example, we expect a 95% CI to cover the truth 95% of the time. In addition to this diagnostic, we apply a calibration procedure described elsewhere [8] to restore nominal coverage by adjusting the CIs. Typically, but not necessarily, the calibrated CI is wider than the nominal CI, reflecting the problems unaccounted for in the standard procedure (such as unmeasured confounding, selection bias, and measurement error)

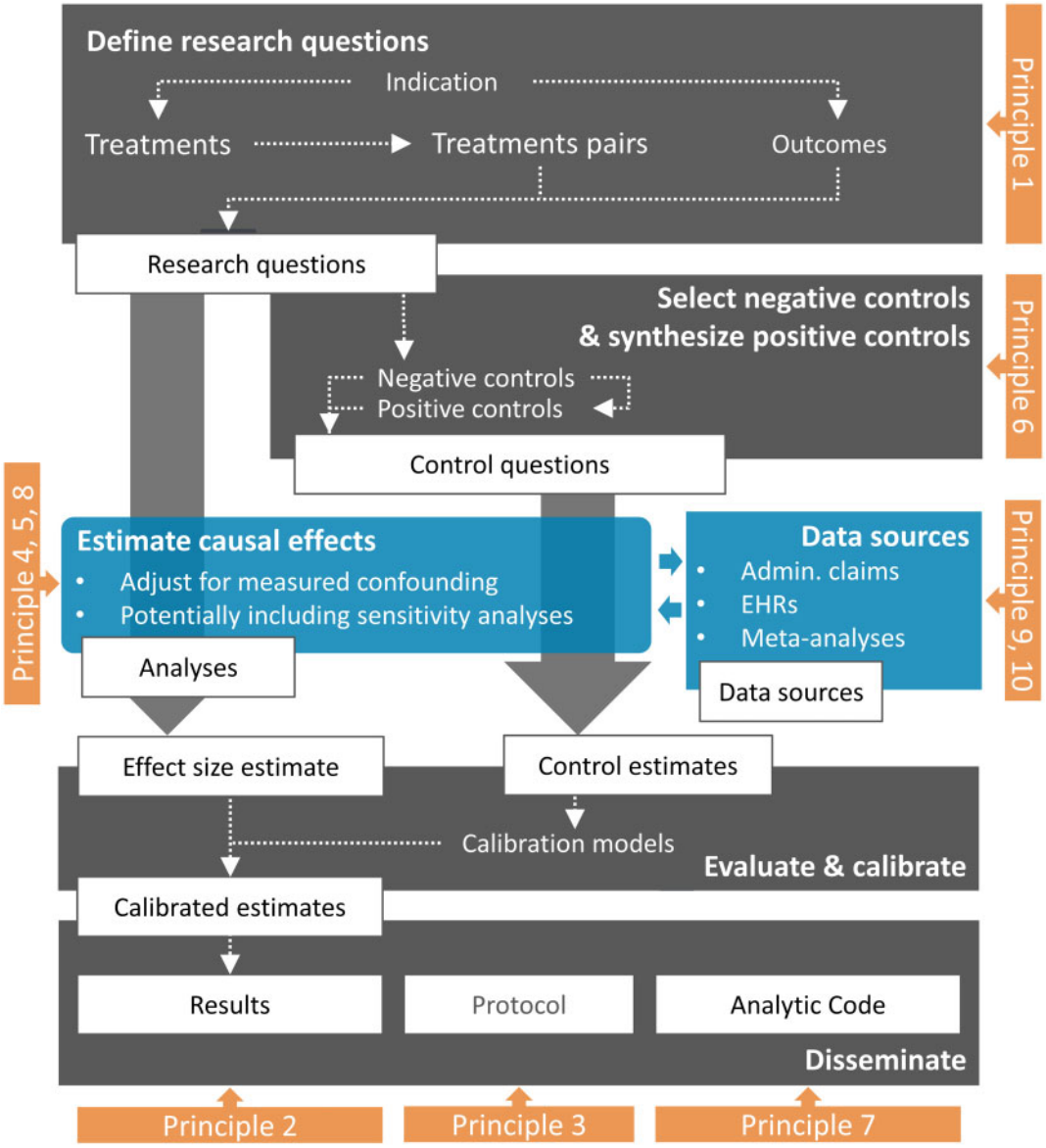


**Figure 1** Overview of a LEGEND study. Admin. Claims: administrative claims; EHRs: electronic health records; LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.

but accounted for in the calibration. A similar process using only negative controls is used to calibrate *P* values.[7]

### Generate the evidence using best practices (Principles 4, 5, and 8)

For each of the research and control questions, we estimate a causal effect size using best practices. For example, we currently employ a new-user cohort design that emulates the randomized experiment that would answer our question of interest,[13] and employ large-scale propensity scores [14] to account for the fact that treatment assignment is not random. Deciding what constitutes "best practices" requires rigorous empirical evaluations; for example, our recent large-scale evaluation of causal effect methods.[15] These evaluations, per Principle 8, fall outside the scope of LEGEND.

To evaluate sensitivity to design choices, it is possible to include various alternative designs; for example, using propensity score matching or stratification.

### Generate the evidence for all questions across a network of databases (Principles 9 and 10)

LEGEND is part of OHDSI, a multi-stakeholder, interdisciplinary collaborative that is striving to bring out the value of observational health data through large-scale analytics. Members of OHDSI have volunteered to participate in LEGEND, agreeing to adhere to the principles and the spirit of collaboration, and this group of collaborators may grow over time. Some LEGEND participants have access to observational data and agree to execute LEGEND studies on their data after acquiring the necessary approvals based on local governance regulations. Each participating data site has translated their data into the Observational Medical Outcomes Partnership Common Data Model (CDM) (https://github.com/OHDSI/CommonDataModel). For each LEGEND study, a study R package is developed and made available as open source (https://github.com/OHDSI/Legend). These study packages implement the entire study, from data in the CDM to the results stored in the results data model described be-

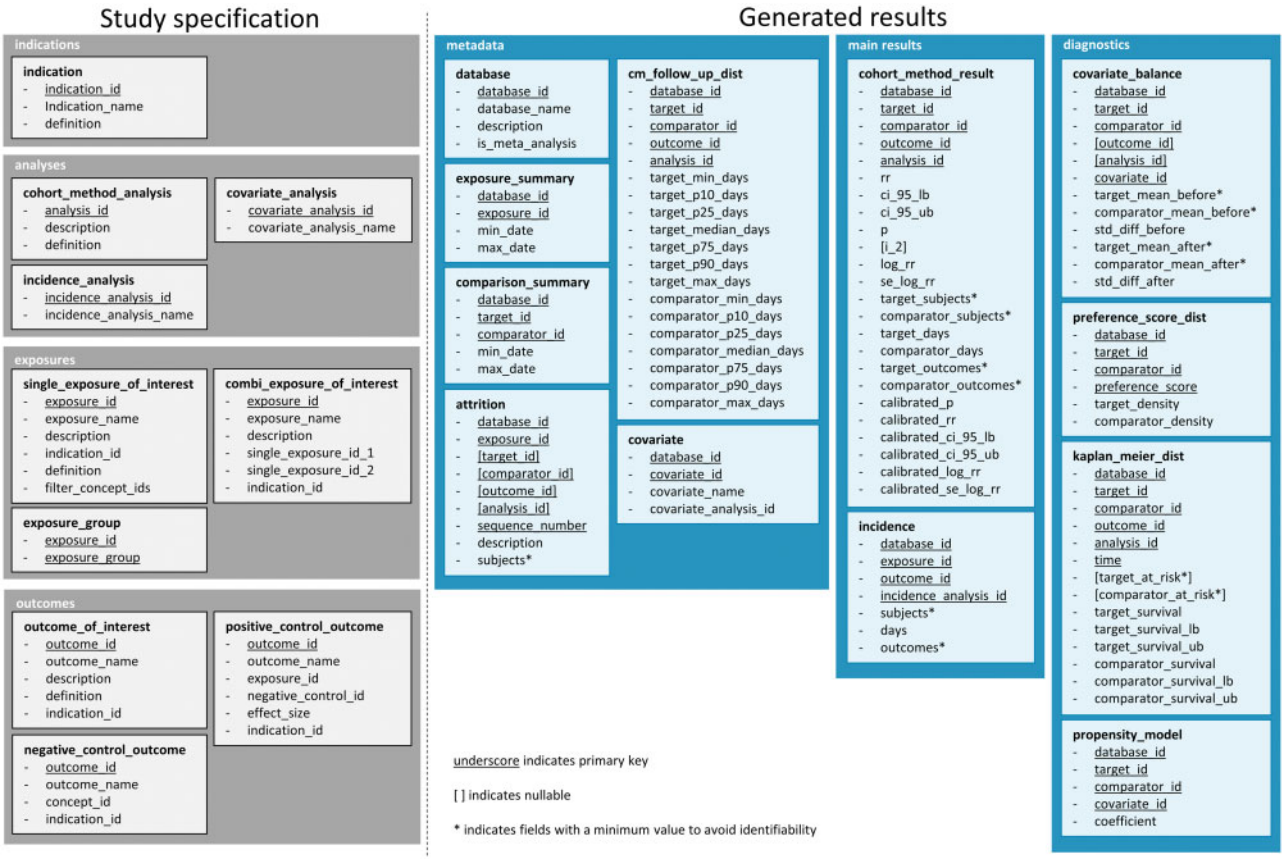


**Figure 2** Data model for storing the LEGEND results, showing the tables and fields per table. LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.

low, including the estimated effect sizes as well as metadata and study diagnostics. These results only include aggregate statistics; no patient-level data are shared. The study packages rely heavily on other open-source software previously developed in OHDSI; specifically, the OHDSI Methods Library (https://ohdsi.github.io/MethodsLibrary/), a set of open-source R packages for performing observational research based on the CDM.

Participating sites are invited to download and install the study package and to execute it against their own data. Results are communicated to the study coordinating center and are synthesized.

### Disseminate the generated evidence (Principle 2)

A LEGEND study is likely to produce a massive number of results, including effect estimates for the many research and control questions, as well as diagnostics, such as covariate balance; additional information, such as population characteristics; and metadata. To manage and communicate that information, we have defined the data model shown in Figure 2 and have described it in detail in the Supplementary Materials.

This data model contains 2 main domains: the first is the full study specifications and the second is the study results. We have developed 2 web applications that connect to the database for exploring the results: the *LEGEND Basic Viewer*, shown in Figure 3, allows users to select a target-comparator-outcome combination and lists all results from the various data sources for that triplet, with drill-down views to understand study population characteristics and diagnostics (http://data.ohdsi.org/LegendBasicViewer/). Additionally, we have created *LEGENDMed Central*, which represents the results as a (virtual) repository of scientific reports, 1 per target-comparator-outcome-database combination. Each report is a PDF that is generated on the fly (http://data.ohdsi.org/LegendMedCentral/). We invite others to develop other applications that promote the dissemination of results using the LEGEND results database.

### Transparency (Principles 3 and 7)

A key guiding principle of the LEGEND approach is transparency. Prior to any analysis, the prespecified LEGEND study protocol and full analytic code are made available in open-source format (https://github.com/OHDSI/Legend).

## DISCUSSION

LEGEND embodies a new approach to generating evidence from health-care data that is designed to overcome weaknesses in the current process of answering and publishing (or not) 1 question at a time. Generating evidence for many research and control questions using a systematic process enables us not only to evaluate that process and the coherence and consistency of the evidence, but also to avoid *P* hacking and publication bias.

The choice of methods used to estimate causal effects should reflect current best practices and is expected to evolve over time as new methods are developed. Even the choice of study design is not cast in stone; for example, depending on the research questions, in

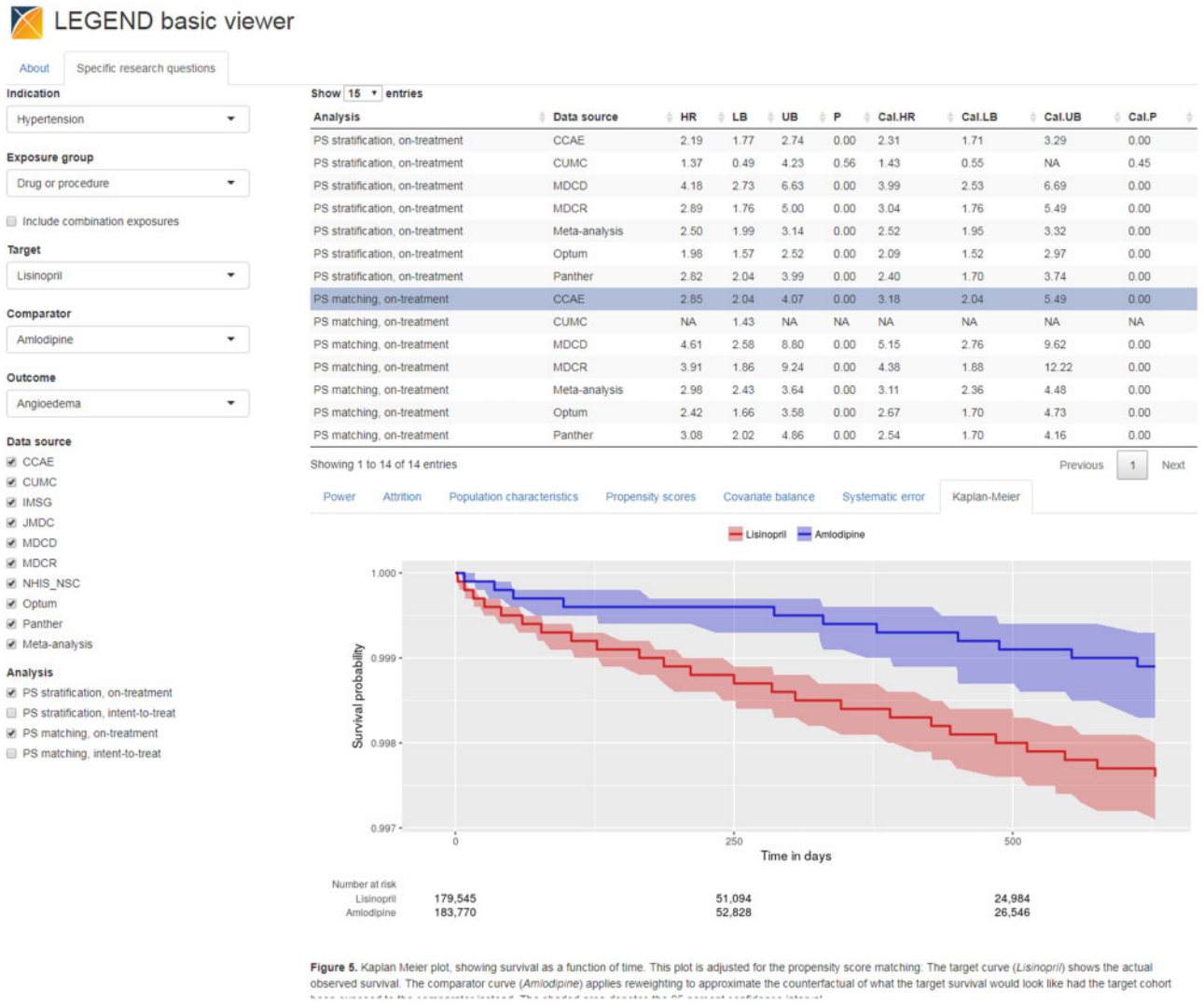


**Figure 3** LEGEND basic viewer: a web-based application for exploring results of the LEGEND hypertension study. LEGEND: Large-scale Evidence Generation and Evaluation across a Network of Databases.

future LEGEND studies we may add self-controlled designs, such as the Self-Controlled Case Series [16], in addition to the current new-user cohort design.

In interpreting LEGEND evidence, researchers must account for multiple hypothesis testing by correcting for however many hypotheses they assess. Paradoxically, performing many analyses helps avoid false positives due to multiple testing because no tests lie hidden, unlike in the current scientific literature, where publication bias is pervasive.[5] Note that empirical calibration by itself does not address multiple testing, but having well-calibrated CIs and *P* values is essential for subsequent adjustments for multiple testing. LEGEND results should always be assessed for quality prior to the consumption of that evidence for medical decision-making. Importantly, the LEGEND framework creates artifacts to assess the internal validity of results, such as covariate balance to assess confounding control, coverage statistics after empirical calibration to assess systematic error, and heterogeneity assessments to assess database consistency. Our framework ensures that such an assessment using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) guidelines[17] is possible. GRADE assessments cover:

1. *The risk of bias*, which we address using best-practice methods by evaluating study diagnostics, such as covariate balance, and by evaluating systematic error through the use of negative and positive controls.
2. *Imprecision*, as expressed in our (calibrated) CIs. By including data from many databases, we typically achieve high precision.
3. *Inconsistency*, which we address through the use of multiple databases and the inspection of between-database heterogeneity.
4. *Indirectness*, through making all possible comparisons.
5. *Publication bias*, through complete dissemination of study results irrespective of the effect estimates.

"Just" generating large amounts of evidence does not guarantee the translation of the evidence generated into better care at the bedside. Although any physician faced with a specific clinical question can directly consult the evidence in the LEGEND results database, the interpretation of that evidence, as discussed above, may prove non-trivial. To bridge the gap between evidence and clinical practice, we suspect an intermediate step must be taken. A form this step can take is papers focused on specific clinical implications, such as our recent paper comparing first-line hypertension treatments at the

class level.[18] It is anticipated that the evidence generated by LEGEND will eventually help to support changes to treatment guidelines, particularly in treatment comparisons for which evidence from randomized trials is unavailable.

## CONCLUSION

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 health-care data, answering many research questions simultaneously using a transparent, reproducible, and systematic approach. Our companion paper demonstrates that the application of LEGEND to antihypertensive treatments produces quality evidence with high internal and external validity. Evidence generated by LEGEND can be used to help inform medical decision-making where evidence is currently lacking.

## FUNDING STATEMENT

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 CONTRIBUTORS

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

None declared.

## REFERENCES

1. Gerstein HC, McMurray J, Holman RR. Real-world studies no substitute for RCTs in establishing efficacy. *Lancet* 2019; 393: 210–1.
2. Rush CJ, Campbell RT, Jhund PS, Petrie MC, McMurray JJV. Association is not causation: treatment effects cannot be estimated from observational data in heart failure. *Eur Heart J* 2018; 39: 3417–38.
3. Ioannidis JPA. Why most published research findings are false. *PLOS Med* 2005; 2(8): e124.
4. 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.
5. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philos Trans A Math Phys Eng Sci* 2018; 376 . doi:10.1098/rsta.2017.0356
6. Schuemie MJ, Ryan PB, Pratt N, *et al*. Large-Scale Evidence Generation and Evaluation across a Network of Databases (LEGEND): Assessing Validity Using Hypertension as a Case Study. *J Am Med Inform Assoc*. 27(8):1268–1277.
7. Schuemie MJ, Ryan PB, DuMouchel W, Suchard MA, Madigan D. Interpreting observational studies: why empirical calibration is needed to correct *P*-values. *Stat Med* 2014; 33: 209–18.
8. Schuemie MJ, Hripcsak G, Ryan PB, Madigan D, Suchard MA. Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proc Natl Acad Sci USA* 2018; 115: 2571–7.
9. Prasad V, Jena AB. Prespecified falsification end points: can they validate true observational associations? *JAMA* 2013; 309: 241–2.
10. Dusetzina SB, Brookhart MA, Maciejewski ML. Control outcomes and exposures for improving internal validity of nonrandomized studies. *Health Serv Res* 2015; 50: 1432–51.
11. Lipsitch M, Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010; 21: 383–8.
12. Voss EA, Boyce RD, Ryan PB, van der Lei J, Rijnbeek PR, Schuemie MJ. Accuracy of an automated knowledge base for identifying drug adverse reactions. *J Biomed Inform* 2017; 66: 72–81.
13. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016; 183: 758–64.
14. Tian Y, Schuemie MJ, Suchard MA. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *Int J Epidemiol* 2018; 47: 2005–14.
15. Schuemie MJ, Soledad Cepede M, Suchard MA, *et al*. How confident are we about observational findings in health care: a benchmark study. 2020; 2 (1). doi: 10.1162/99608f92.147cc28e
16. Whitaker HJ, Farrington CP, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med* 2006; 25: 1768–97.
17. Guyatt GH, Oxman AD, Kunz R, *et al*. What is 'quality of evidence' and why is it important to clinicians? *BMJ* 2008; 336: 995–8.
18. 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.