Journal of the American Medical Informatics Association, 00(0), 2023, 1–10
https://doi.org/10.1093/jamia/ocad009
Research and Applications

Research and Applications

Reproducible variability: assessing investigator discordance across 9 research teams attempting to reproduce the same observational study

Anna Ostropolets[1], Yasser Albogami[2], Mitchell Conover[3], Juan M. Banda[4], William A. Baumgartner Jr[5], Clair Blacketer[3], Priyamvada Desai[6], Scott L. DuVall[7,8], Stephen Fortin[3], James P. Gilbert[3], Asieh Golozar[9], Joshua Ide[10], Andrew S. Kanter[1], David M. Kern[3], Chungsoo Kim[11], Lana Y.H. Lai[12], Chenyu Li[13], Feifan Liu[14], Kristine E. Lynch[7,8], Evan Minty[15], Maria Inês Neves[16], Ding Quan Ng[17], Tontel Obene[18], Victor Pera[19], Nicole Pratt[20], Gowtham Rao[3], Nadav Rappoport[21], Ines Reinecke[22], Paola Saroufim[23], Azza Shoaibi[3], Katherine Simon[24], Marc A. Suchard[25,26], Joel N. Swerdel[3], Erica A. Voss[3], James Weaver[3], Linying Zhang[1], George Hripcsak[1,27], and Patrick B. Ryan[1,3]

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, New York, USA, [2]Department of Clinical Pharmacy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia, [3]Observational Health Data Analytics, Janssen Research & Development, Titusville, New Jersey, USA, [4]Department of Computer Science, Georgia State University, Atlanta, Georgia, USA, [5]Division of General Internal Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA, [6]Research IT, Technology and Digital Solutions, Stanford Medicine, Stanford, California, USA, [7]VA Salt Lake City Health Care System, Salt Lake City, Utah, USA, [8]Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, USA, [9]Odysseus Data Services, New York, New York, USA, [10]Johnson & Johnson, Titusville, New Jersey, USA, [11]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea, [12]Department of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, UK, [13]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, [14]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, Massachusetts, USA, [15]O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Calgary, Canada, [16]Real World Solutions, IQVIA, Durham, North Carolina, USA, [17]Department of Pharmaceutical Sciences, School of Pharmacy & Pharmaceutical Sciences, University of California, Irvine, California, USA, [18]Mississippi Urban Research Center, Jackson State University, Jackson, Mississippi, USA, [19]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands, [20]Quality Use of Medicines and Pharmacy Research Centre, University of South Australia, Adelaide, Australia, [21]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Israel, [22]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany, [23]Cleveland Institute for Computational Biology, Case Western Reserve University, Cleveland, Ohio, USA, [24]VA Tennessee Valley Health Care System, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [25]Department of Biostatistics, University of California, Los Angeles, California, USA, [26]Department of Human Genetics, University of California, Los Angeles, California, USA, and [27]Medical Informatics Services, New York-Presbyterian Hospital, New York, New York, USA

Corresponding Author: Patrick B. Ryan, Epidemiology Analytics, Janssen Research & Development, 1125 Trenton Harbourton Road, Titusville, NJ 08560, USA; ryan@ohdsi.org

Received 25 November 2022; Revised 4 January 2023; Editorial Decision 17 January 2023; Accepted 23 January 2023

ABSTRACT

Objective: Observational studies can impact patient care but must be robust and reproducible. Nonreproducibility is primarily caused by unclear reporting of design choices and analytic procedures. This study aimed to: (1) assess how the study logic described in an observational study could be interpreted by independent researchers and (2) quantify the impact of interpretations' variability on patient characteristics.

Materials and Methods: Nine teams of highly qualified researchers reproduced a cohort from a study by Albogami et al. The teams were provided the clinical codes and access to the tools to create cohort definitions such that the only variable part was their logic choices. We executed teams' cohort definitions against the database and compared the number of subjects, patient overlap, and patient characteristics.

Results: On average, the teams' interpretations fully aligned with the master implementation in 4 out of 10 inclusion criteria with at least 4 deviations per team. Cohorts' size varied from one-third of the master cohort size to 10 times the cohort size (2159–63 619 subjects compared to 6196 subjects). Median agreement was 9.4% (interquartile range 15.3–16.2%). The teams' cohorts significantly differed from the master implementation by at least 2 baseline characteristics, and most of the teams differed by at least 5.

Conclusions: Independent research teams attempting to reproduce the study based on its free-text description alone produce different implementations that vary in the population size and composition. Sharing analytical code supported by a common data model and open-source tools allows reproducing a study unambiguously thereby preserving initial design choices.

Key words: reproducibility, observational data, credibility, open science

INTRODUCTION

Observational studies conducted on electronic health record and administrative claims data have the potential to impact decision making, especially in cases where randomized clinical trials are not readily available or feasible.[1] However, uncertainty about the ability of analytical methods to mitigate bias and uncertainty about internal validity as well as nontransparency in reporting of methods and results contribute to the concerns raised about credibility of observational evidence.[2]

Reproducibility of findings has been commonly viewed as a means of improving reliability and robustness of studies thereby building trust in their results.[3] Reliable evidence should be reproducible such that a different researcher should be able to perform the same task of executing a given analysis on a given dataset and produce an identical result as another researcher. In the context of retrospective analysis of observational healthcare databases, reproducibility requires the process to be fully specified, generally in both human-readable and computer-executable form, such that no study implementation decisions are left to the discretion of the investigator.[4]

There is a concern that clinical informatics may also face the "reproducibility crisis" that has been observed across multiple scientific disciplines.[5,6] Previous studies have shown that reproducing clinical studies requires involvement of the original author(s) of the study and, even after such, a quarter of the studies were not fully reproducible.[7] In preclinical and psychological studies, less than half of the reproduced findings had the same direction and statistical significance.[8,9] In observational database research, Wang et al[4] showed that only half of the point estimates and confidence intervals of the reproduced studies had the same direction (were on the same side of null) as the original implementations. These studies demonstrate that having open data or access to the same data source was insufficient for reproducibility and that nonreproducibility was primarily caused by unclear reporting of design choices and analytic procedures. Therefore, a critical challenge in reproducibility is ambiguity and the lack of specificity associated with natural language description of study design. In the absence of source code to fully repeat an analysis and a data source that was preprocessed exactly as the original data source, investigator-induced error may occur with the interpretation and translation of natural language descriptions into a new implementation.

Several initiatives and working groups have been established to improve study design reporting, namely arguing for the field to improve its prespecification of analyses through public registration and posting of protocols and detailed reporting of methodological design decisions in manuscripts. Similar to the templates developed for study protocols, several templates and checklists, such as STROBE, RECORD, ENCePP, and TRIPOD guidelines, have been proposed for study reporting.[10–14] While promising, it is unclear whether detailed reporting by the original investigators can be consumed by independent research groups and consistently interpreted and reimplemented. Given that practices for study protocol prespecification remain severely underutilized, it is also not clear to what extent such templates will be consistently and comprehensively used by the broader research community.[5,15,16]

In this work, we aimed to: (1) assess how the study design described in an observational research study could be interpreted by multiple teams of independent researchers and (2) quantify the impact of the variability of replication design choices made by those teams on patient characteristics.

Reproducing a study entails several design steps: choosing the hypothesis; defining entities like the treatments, outcomes, and patient characteristics as sets of codes (concept sets); defining the logic around those concept sets to produce phenotype algorithms (cohort definitions) that, when executed against a data source, translate into cohorts of study subjects; statistical analyses; diagnostics and sensitivity analyses; and presentation of results. This study addressed a single step, the definition of the logic of a phenotype algorithm while preserving the other elements such as concept sets stable.

MATERIALS AND METHODS

This study included selecting a paper for replication, constructing and executing a cohort definition from the paper together with the

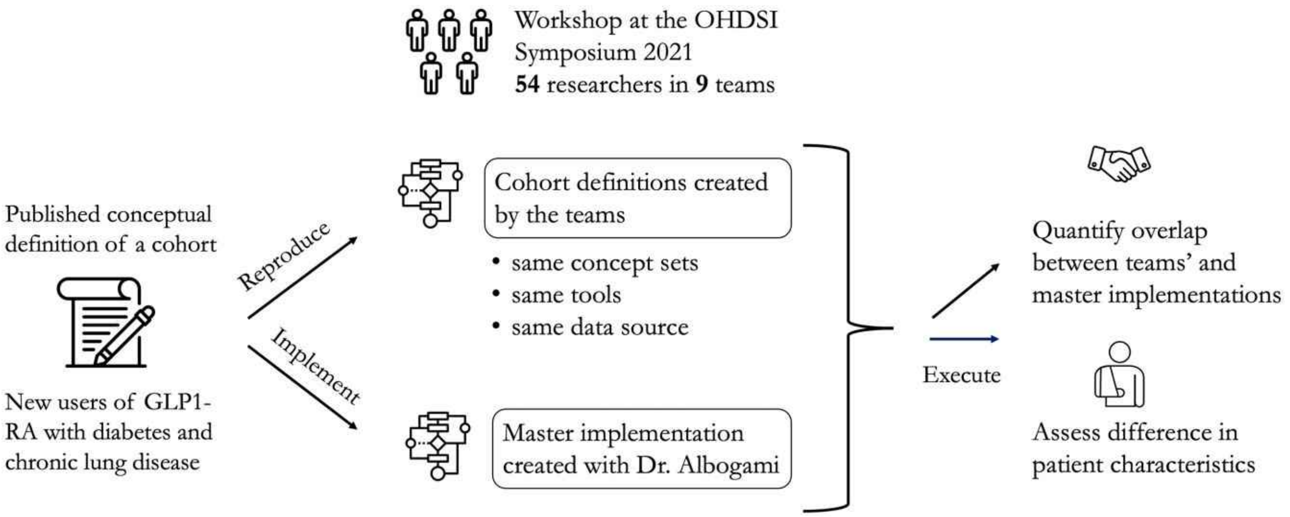
Workshop at the OHDSI Symposium 2021
54 researchers in **9** teams
Published conceptual definition of a cohort
New users of GLP1-RA with diabetes and chronic lung disease
Reproduce
Implement
Cohort definitions created by the teams
• same concept sets
• same tools
• same data source
Master implementation created with Dr. Albogami
Execute
Quantify overlap between teams' and master implementations
Assess difference in patient characteristics

Figure 1. Study design overview.

original author, having 9 teams independently implement the cohort definition, executing their definitions on the same data source and subsequently comparing patient selection and composition. Overview of the study is presented in Figure 1.

Publication for reproduction

To select the study for reproduction, we screened PubMed articles published after 2020. We then reviewed the papers for sufficient description of the study methods using the text in the main body, in diagrams and attrition tables and Supplementary Materials. Another criterion was having the same data source at our disposal for reproduction. We selected the article by Albogami et al[17] based on the availability of the data source used and the completeness of the study design description in the main body of the text and Supplemental Materials. The study was published in 2021 and investigated an association of glucagon-like peptide 1 receptor agonists (GLP-1RA) and chronic lower respiratory disease (CLRD) exacerbation in a population with type 2 diabetes mellitus (T2D) and CLRD. The study suggested a strong negative association between GLP1-RA use and CLRD exacerbations (hazard ratio 0.52, 95% confidence interval [CI] 0.32–0.85 for inpatient CLRD admissions and incidence rate ratio 0.70, 95% CI 0.57–0.87 for outpatient CLRD visits). The study used sound statistical approaches to mitigate potential bias: inverse probability of treatment weights to adjust for confounding and a number of sensitivity analyses with different statistical methods (Bayesian additive regression trees, propensity score matching, imputation techniques for obesity and tobacco dependence, and negative control outcomes) and design choices (using sulfonylureas as an alternative comparator, supplementing principal position in claims for CLRD diagnosis and using GLP1-RA drugs as an add-on to metformin). The authors extensively described the methods in the paper, supplied the codes for T2D and CLRD, provided a flow chart for patient selection as well as for study timeline, inclusion and exclusion criteria using the diagram previously suggested for reporting.[18]

Conceptual definition

For the purpose of this manuscript, we focused on the target patient cohort, which was defined in the manuscript as new users of GLP1-RA add-on therapy aged more than 17 years with at least 1 outpatient or 2 inpatient encounters with T2D and CLRD in the year before the index date with no prior insulin or dipeptidyl peptidase 4 inhibitors exposure and no prior type 1 diabetes mellitus, cystic fibrosis, lung cancer, pulmonary embolism, pulmonary hypertension, conditions requiring chronic systemic corticosteroid therapy within a year or pregnancy at the index date.

Master implementation

Based on the conceptual definition of the cohort, together with the original author, we constructed a target cohort definition using the Observational Health Data Sciences and Informatics (OHDSI) tool ATLAS (Figure 2). ATLAS is a web-based application that allows defining phenotypes, constructing and executing cohorts against local data source(s), characterizing subjects in a cohort, and designing and implementing various observational studies.[19] The definition specified the entry event upon which a patient enters the cohort (first GLP1-RA exposure in 2007–2017), 10 inclusion and exclusion criteria and the exit event upon which the patient leaves the cohort (is right-censored). Each inclusion and exclusion criterion comprised a start and end date, a duration (for drug exposures), one or multiple associated concept sets, a set of Boolean or temporal logic applied to the concept set(s) and an order in which the criteria were applied. The master implementation used a list of predefined concept sets created in collaboration with the original author (Supplementary Table S1).

When creating the master implementation, we also assessed the influence of each individual criterion on patient selection when executed against the same data source used in the original study (IBM® MarketScan® Commercial database, Table 1). Several criteria, such as not being pregnant on the index date or being older than 17, had negligible impact on patient selection as subjects with T2D are likely to be older and, therefore, not pregnant. The requirement of the first GLP1-RA exposure within 365 days did not have large influence on patient attrition because we initially chose the earliest event in the cohort.

On the other hand, requiring a prior CLRD diagnosis, at least a year of prior observation and add-on antidiabetic therapy had a large impact on patient selection with only 7.8%, 55.3%, and 56.2% of subjects satisfying these criteria, respectively. Requiring no prior insulin exposure eliminated some subjects, but the influence

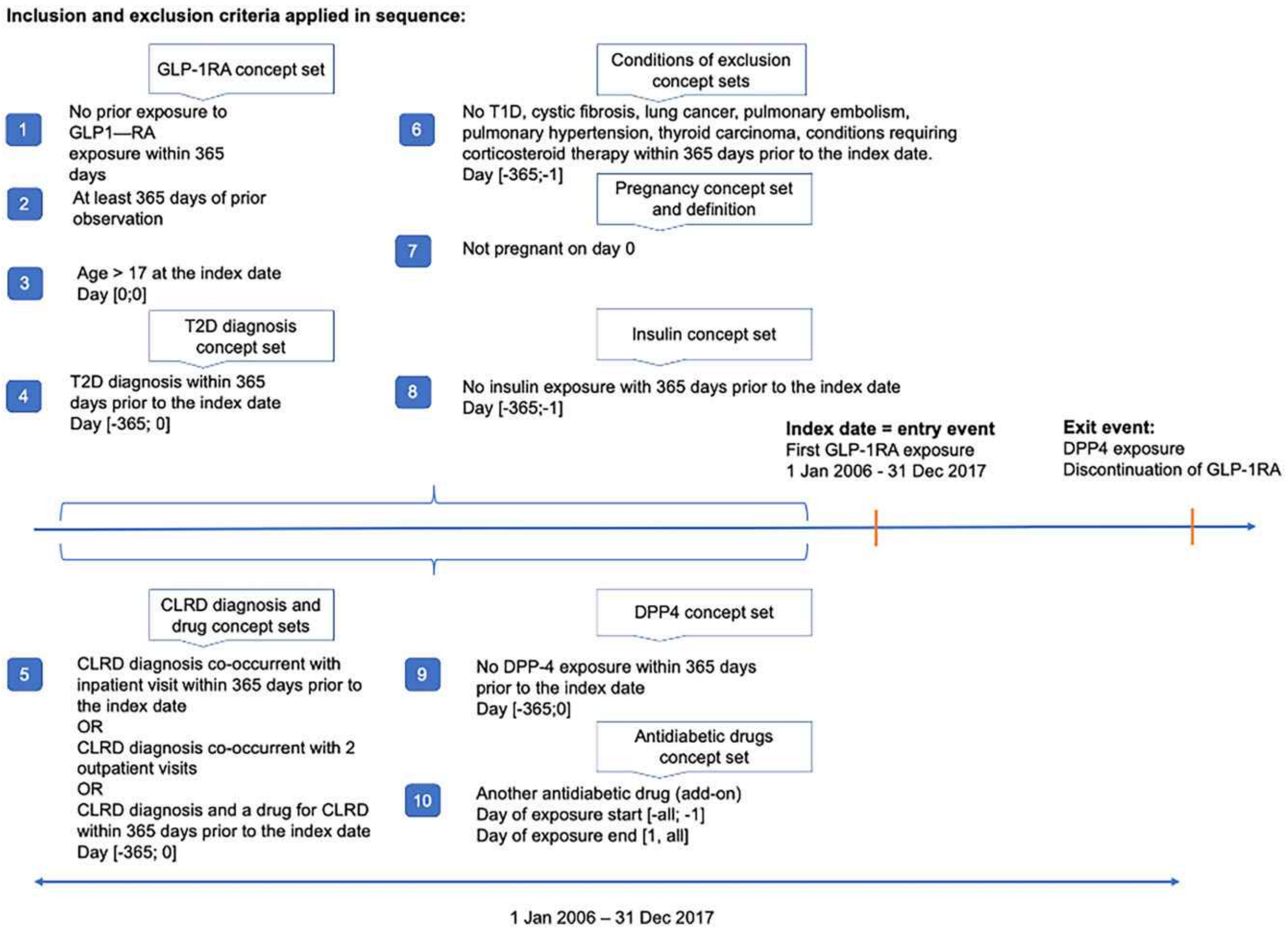
Inclusion and exclusion criteria applied in sequence:
GLP-1RA concept set
1
No prior exposure to GLP1—RA exposure within 365 days
2
At least 365 days of prior observation
3
Age > 17 at the index date
Day [0;0]
T2D diagnosis concept set
4
T2D diagnosis within 365 days prior to the index date
Day [-365; 0]
Conditions of exclusion concept sets
6
No T1D, cystic fibrosis, lung cancer, pulmonary embolism, pulmonary hypertension, thyroid carcinoma, conditions requiring corticosteroid therapy within 365 days prior to the index date.
Day [-365;-1]
Pregnancy concept set and definition
7
Not pregnant on day 0
Insulin concept set
8
No insulin exposure with 365 days prior to the index date
Day [-365;-1]
Index date = entry event
First GLP-1RA exposure
1 Jan 2006 - 31 Dec 2017
Exit event:
DPP4 exposure
Discontinuation of GLP-1RA
CLRD diagnosis and drug concept sets
5
CLRD diagnosis co-occurrent with inpatient visit within 365 days prior to the index date
OR
CLRD diagnosis co-occurrent with 2 outpatient visits
OR
CLRD diagnosis and a drug for CLRD within 365 days prior to the index date
Day [-365; 0]
DPP4 concept set
9
No DPP-4 exposure within 365 days prior to the index date
Day [-365;0]
Antidiabetic drugs concept set
10
Another antidiabetic drug (add-on)
Day of exposure start [-all; -1]
Day of exposure end [1, all]
1 Jan 2006 – 31 Dec 2017

Figure 2. Master new glucagon-like peptide 1 receptor agonists (GLP1-RA) user cohort implementation: entry and exit event and 10 inclusion and exclusion criteria.

of this criteria was limited by the fact that we excluded GLP1-RA and insulin combinations from the list of target drugs for selecting the target subjects.

Study settings

Reproduction of the master cohorts was organized as a 1-day workshop, which was held as a part of Observational Health Data Sciences and Informatics (OHDSI) 2021 Global Symposium on September 13, 2021. OHDSI is an international multistakeholder, interdisciplinary data network of electronic health records, administrative claims, hospital discharge data, registries, and other observational data sources that standardizes data through a common data model (Observational Medical Outcomes Partnership Common Data Model, OMOP CDM) and harmonizes the content of the CDM through applying a common reference vocabulary system (OMOP Standardized Vocabularies). OHDSI encompasses more than 800 million patients across 30 countries.

Prior to the Symposium, we invited all OHDSI collaborators to participate in the challenge. Fifty-four collaborators met all prerequisites (familiarity with the paper, OMOP CDM, Standardized Vocabularies, and OHDSI tools) and were divided into 9 groups supervised by 2 workshop co-hosts (PBR and AO).

Each group had at least one informatician with extensive CDM and ATLAS knowledge and one epidemiologist or clinical expert. They were provided with access to an ATLAS instance with an empty cohort definition template. The ATLAS instance was prepopulated with the same predefined concept sets used in the master implementation such that the reproducibility exercise was focused on the logic of the cohort definitions and not on selection of the correct drug and diagnosis codes which in itself is challenging and would introduce significant variation. Over the day, each team separately implemented the cohort definition based on their interpretation of the paper and the Supplementary Materials. Groups could define any number of criteria in their implementation and apply them in any order.

Data analysis

All cohort definitions were subsequently executed on the IBM® MarketScan® Commercial database (CCAE) and compared to the master implementation of the cohort definitions created together with the original author. For each cohort, the number of subjects and demographic characteristics (age and sex) at index date were extracted, along with diseases and drugs used as recorded in the 365 days prior to the index date. To assess the influence of the design choices on patient selection, we calculated the agreement between each cohort created by the participants and the master cohort using the Jaccard index[20] defined as the number of subjects included in both cohorts divided by the total number of subjects in

Table 1. Criteria used to define master implementation and the number of subjects satisfying each individual criterion in the cohort executed against IBM® MarketScan® Commercial database

	Criteria	Subjects who satisfied the criteria, *n* (%)
Cohort entry	First glucagon-like peptide 1 receptor agonists (GLP1-RA) exposure in 2005–2017	570 664 (100)
1	Had no GLP1-RA exposure within 365 days prior to the index date	563 245 (98.7)
2	Had at least 365 days of prior observation	315 616 (55.3)
3	Age >17	569 757 (99.9)
4	Had type 2 diabetes mellitus (T2D) within 365 days prior to the index date	430 080 (75.4)
5	Had chronic lower respiratory disorder (CLRD) within 365 days prior to or on the index date	44 668 (7.8)
6	Had no type 1 diabetes, cystic fibrosis, lung cancer, pulmonary embolism, pulmonary hypertension, thyroid carcinoma, conditions requiring corticosteroid therapy within 365 days prior to the index date	488 606 (85.7)
7	Was not pregnant on the index date	565 877 (99.2)
8	Had no insulin exposure within 365 days prior to or on the index date	402 407 (70.5)
9	Had no dipeptidyl peptidase-4 (DPP4) inhibitor exposure within 365 days prior to or on the index date	474 365 (83.2)
10	Had another T2D drug that started before the index date and ended on or after the index date	320 658 (56.2)
	All criteria	6196 (1.1)

either cohort. Additionally, we extracted the variables used to describe the population in the original study and calculated the standardized difference of means between each cohort and the master implementation for each variable.[21]

RESULTS

Comparison of the master implementation and each team's implementations

On average, each team's interpretation fully aligned with the master implementation in 4 out of 10 inclusion criteria; all teams had at least 4 criteria deviating from the master implementation. As shown in Figure 3, all 9 teams fully reproduced 2 criteria: (1) having 365 days of prior observation; and (2) age greater than 17 years at the index date. Two additional criteria were implemented correctly by the majority of the 9 teams: (3) no conditions of exclusion within 365 days prior (1 of 9 teams implemented this differently), and (4) no insulin exposure (4 of 9 teams implemented this differently).

Other criteria were less reproducible. Interpretation of the criteria requiring complex logic was highly variable. As per the paper, the subjects had to have "… at least one inpatient or two outpatient encounters with T2D and CLRD, defined based on the presence of diagnoses or medication dispensing…during the year before index date." This criterion was implemented as (1 inpatient visit with CLRD diagnosis OR 2 outpatient visits with CLRD diagnosis OR CLRD drug exposure and CLRD diagnosis within −365 to 0 day prior to the index date) AND (1 T2D diagnosis within −365 to 0 day prior to the index date).

Subsequently, the teams implemented it as follows:

a. (1 inpatient visit with CLRD diagnosis OR 2 outpatient visits with CLRD diagnosis OR CLRD drug exposure within −365 to −1 day prior to the index date) AND (1 inpatient visit with T2D diagnosis OR 2 outpatient visits with T2D diagnosis OR T2D drug exposure within −365 to −1 day prior to the index date);
b. (1 inpatient visit with (CLRD diagnosis OR CLRD drug exposure) OR 2 outpatient visits with (CLRD diagnosis OR CLRD drug exposure) within −365 to −1 day prior to the index date) AND (1 inpatient visit with (T2D diagnosis OR T2D drug exposure) OR 2 outpatient visits with (T2D diagnosis OR T2D drug exposure) within −365 to −1 day prior to the index date);
c. (1 inpatient visit with (CLRD diagnosis OR CLRD drug exposure) OR 2 outpatient visits with (CLRD diagnosis OR CLRD drug exposure) within −365 to 0 day prior to the index date) AND (1 inpatient visit with (T2D diagnosis OR T2D drug exposure) OR 2 outpatient visits with (T2D diagnosis OR T2D drug exposure) within −365 to 0 day prior to the index date);
d. ((1 inpatient visit with (T2D drug exposure OR T2D diagnosis) AND (CLRD drug exposure OR CRLD diagnosis)) OR (2 outpatient visit with (T2D drug exposure OR T2D diagnosis) AND (CLRD drug exposure OR CRLD diagnosis)) within −365 to −1 day prior to the index date).

Here, variation in implementation stemmed from different combinations of timing of events, their co-occurrence, and combination of individual subcriteria.

Similarly, the criterion of add-on therapy was implemented in 3 different ways: (a) having another antidiabetic drug on the index date, (b) having an overlapping drug exposure that starts before the index date and ends after the index date, and (c) having drug exposure with a typical number of days of supply.

A detailed description of the deviations per each criterion is provided in Supplementary Table S2.

Influence of different choices on patient characteristics

We observed high variation in cohort size from having one-third of the master implementation patient count to having 10 times the cohort size (2159–63 619 subjects compared to 6196 subjects in the master implementation). Not surprisingly, the agreement between the master cohort and the teams' implementations also varied greatly (Figure 4).

Median agreement was 9.4% (interquartile range 15.3–16.2%) and ranged between 0% in Team 5 and 35.4% in Team 8. Similarly,

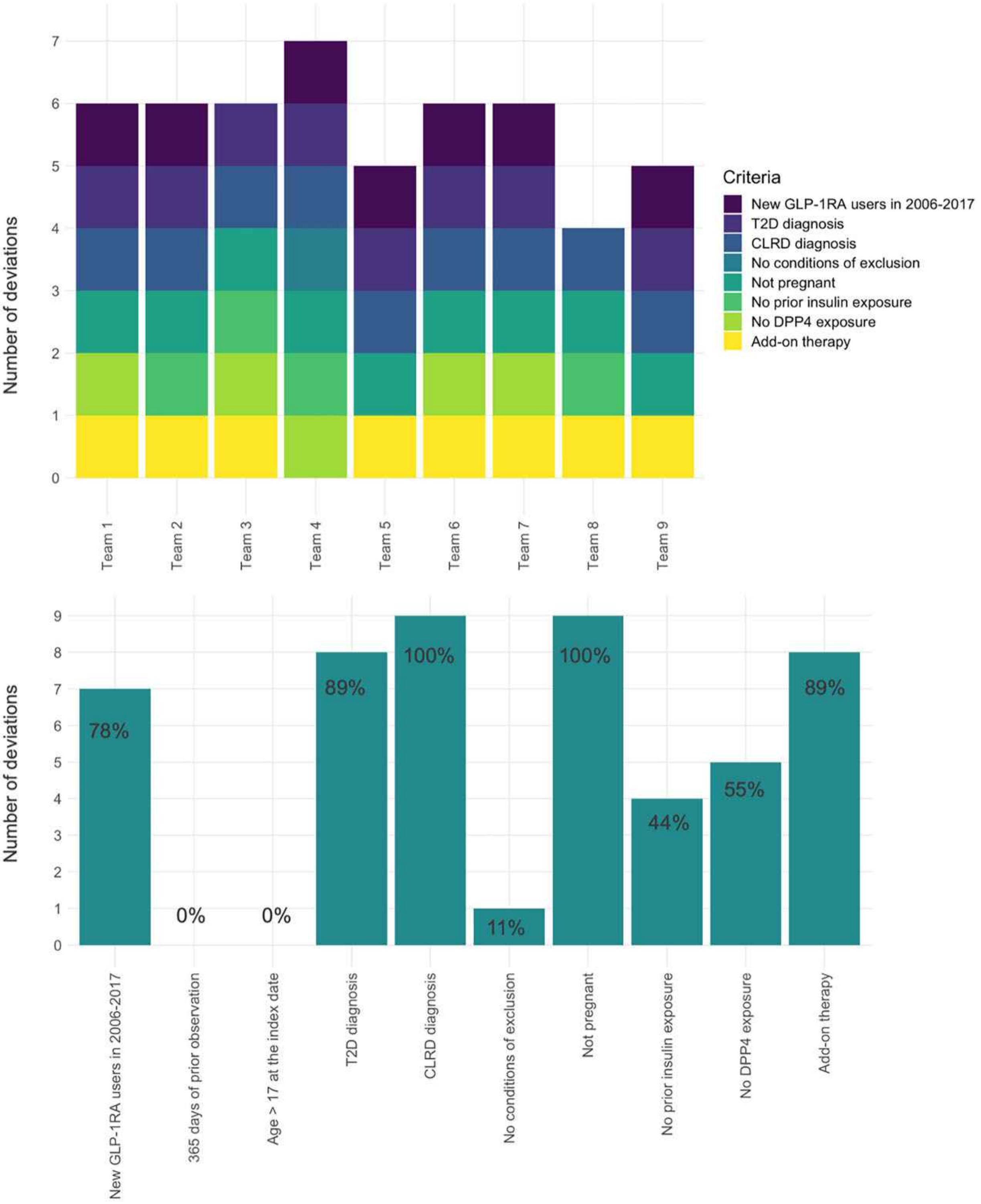

Figure 3. Number of deviations per inclusion and exclusion criteria and team.

the teams' implementations differed from each other greatly (median agreement was 10.0% and interquartile range 0.0–17.5%).

Patient characteristics

The age distribution was similar across all cohorts with 45–64 years old being the major age group (Supplementary Table S3). The gender distribution was also similar to the master implementation except for cohorts of teams 4 and 5 that had a lower proportion of females (58.3% and 57.4% compared to 66.2% in master).

As shown in Figure 5, the cohort generated from each team's implementation differed from the master implementation by at least 2 baseline characteristics with a standardized difference of means (SDM) >0.1, and the majority of the teams differed by at least 5

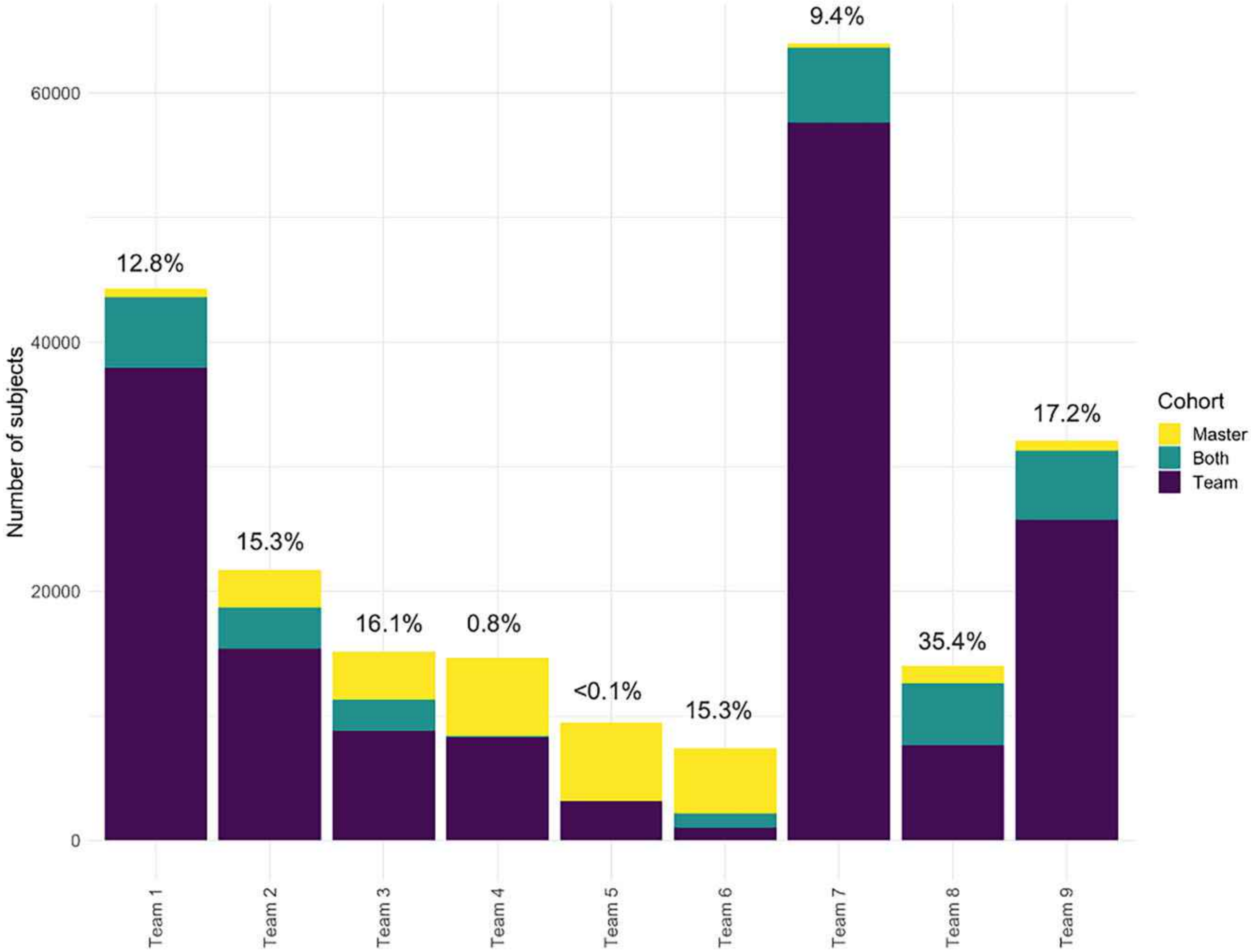

Figure 4. Cohort overlap for each team's cohort and the master implementation, number of subjects, and agreement (Jaccard index).

baseline characteristics. The difference was especially prominent for chronic lung disease disorder, asthma, and prior metformin exposure, which corresponded to the largest number of deviations in implementing those criteria. Cohorts were generally similar in prevalence of conditions related to T2D such as glaucoma or hypoglycemia.

DISCUSSION

This study demonstrates that using natural language to describe complex study design logic produces high variability in interpretation. We showed that 9 teams of qualified investigators, given the exact same task of reproducing a study cohort using consistent development tools and predefined concept sets, obtained 9 different cohort definitions with 52 deviations in total across a set of 10 inclusion and exclusion criteria. In this experiment, we sought to reproduce demonstrated current best practices in observational studies. It provided detailed descriptions, including following many of the recommendations in design reporting[13] such as creating an attrition diagram of sequential inclusion criteria and drawing a figure to show covariate temporal windows.

Nevertheless, as demonstrated by the examples of T2D and CLRD, add-on therapy and pregnancy, natural language does not allow representing such complex logic and co-occurring events in an unambiguous way. All eligibility criteria should be accompanied by explicit code sets as the code set construction process itself is a highly variable process and introduces significant variability.[22] Even if the code sets are provided, some criteria are phenotypes in their own right and require further detailed specification. For example, as opposed to chronic conditions like diabetes mellitus that are believed to be lifelong and therefore could be modeled by looking for the existence of a diagnosis any time in a patient's medical history, pregnancy is a temporal state that has a clear start and end date which need to be accurately specified to capture a pregnancy episode and determine where those episode duration occur in relation to the index date. Given complexity of this task, there are numerous standalone papers that focus on algorithms for pregnancy, varying from simple algorithms to more complex definitions of separate pregnancy episodes.[23–29] Merely providing codes, therefore, is not sufficient to fully specify an inclusion or exclusion criteria.

Even simple concepts like age can be applied at different points in the algorithm and can be represented as age ≥18 years or age >17 years. While in these cases variations do not lead to changes in the patient cohort, more complex criteria influence patient selection to a great extent.

While adding graphs and figures facilitates interpretation of the criteria, the heterogeneity in implementation by our teams suggests

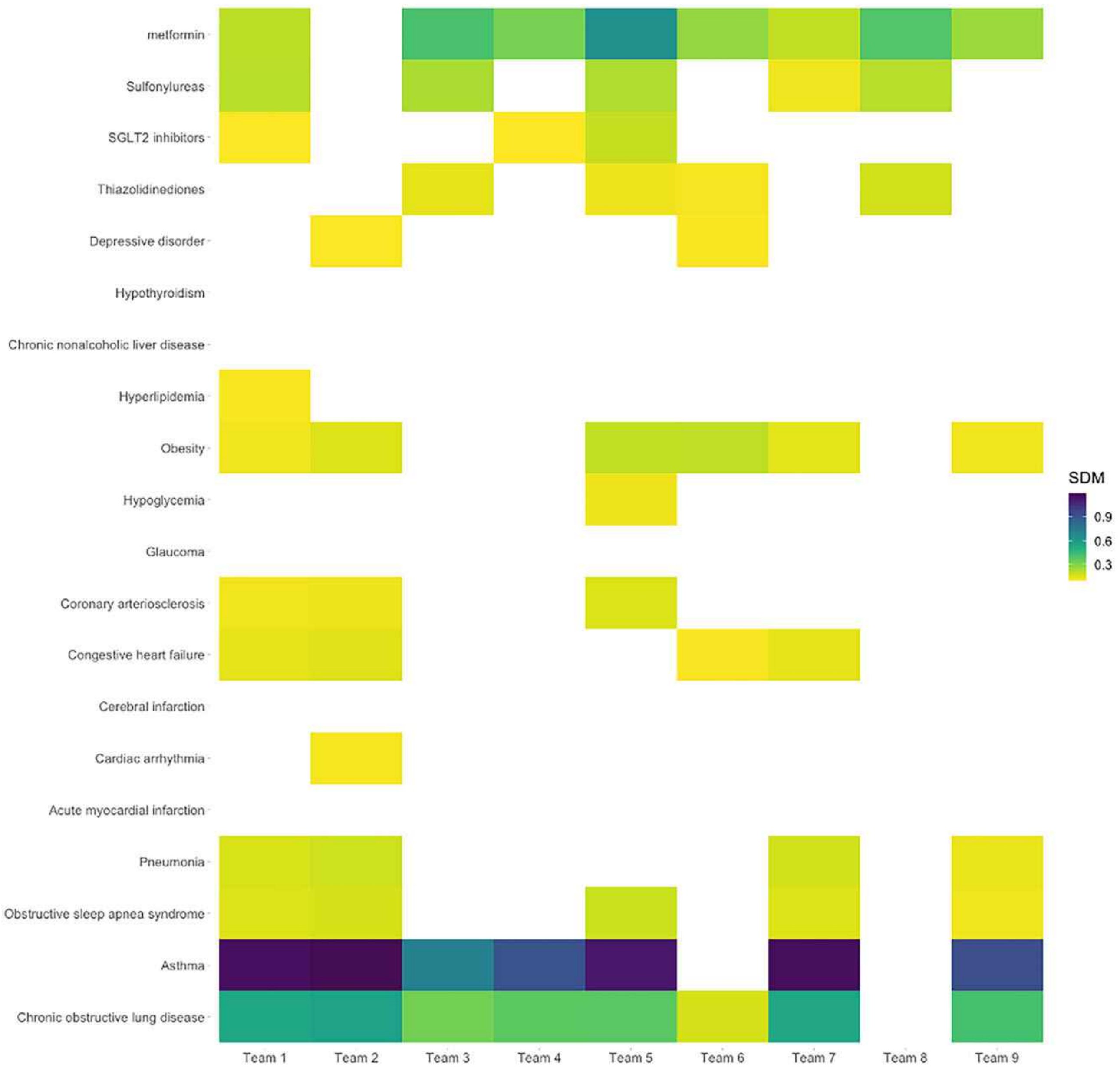

Figure 5. Difference in patient characteristics between the master implementation and teams' implementations colored based on the absolute standardized difference of means (SDM). White indicates SDM < 0.1.

that the problem rests not with the skill of the investigators but with the process for generating and disseminating evidence itself. We were able to compare the implementations to the master implementation due to the data source availability. Nevertheless, we still observed differences between the patient characteristics in the master cohort compared to the patient characteristics in the Albogami paper, which may partly be due to the differences in the cut of CCAE supplied to different institutions. It highlights the fact that reproducing phenotyping logic requires knowing the underlying data schema so the logic can be applied to the proper tables, columns, and data elements. Otherwise, the step of inferring logic must be accompanied by inferring how logic is imposed on the data elements. Having a common data model removes this inference step and directly reproduces the logic on data that have been standardized to a common format.

In the OHDSI community, analysis packages are developed to have a consistently defined input of an observational database formatted into the OMOP CDM. In this exercise, by using the OHDSI ATLAS tool, the analysis was fully specified as a cohort definition with a defined output: a cohort table populated with all persons satisfying criteria for a duration of time, and summary results about the cohort, including cohort count and attrition from each inclusion criteria. The cohort definitions can be exported as analytical code, which can be directly implemented on any data source that is mapped to OMOP CDM without the need to interpret it and without the risk of introducing bias while implementing the interpretation.

Free text study description can serve as a helpful guide to gain an understanding of conceptual study design, but the analysis source code should be considered the referent "gold standard" of truth of what was computed.[30,31] Natural language description can be generated from the code in an automated fashion. On the other hand, code cannot be generated reliably based on the description, which may be inconsistent, ambiguous, or missing sufficient detail in the paper, diagrams, templates, and Supplementary Materials. Extensive and detailed specification of study design and all parameters and elements may not be possible due to editorial constraints and word limit or may be overwhelming for a reader. Other solutions to improve reproducibility (such as using pseudocode or filling prespecified templates and checklists) can be placed as Supplementary Materials thus avoiding word limit and improving paper readability, but, as demonstrated before, lack traction in the broader scientific community.[15,32]

In this regard, if we truly aspire to reproducible science, we should not hope that good documentation is sufficient and tolerate optional sharing of code, but rather make code sharing a hard requirement that can be complemented by free text descriptions.[33,34]

There were limitations to the experiment. While the teams were introduced to the study before the workshop and found a full day to be sufficient to discuss and reproduce the cohorts, the activity was limited to 8 h. Therefore, it is possible that time constraints influenced study findings. To improve the generalizability and robustness of these results, future work may involve repeating this style of experiment with a larger number of studies over a prolonged period of time. We ensured that all teams had at least one clinician, bioinformatician, and a team member who was familiar with the data and tools, but individual level of expertise may have varied. We selected one study as it was not feasible to have multiple teams perform multiple studies, but it is possible that the experience with this study may not be generalizable to other studies.

CONCLUSIONS

Reproducibility of observational studies is currently limited by lack of sharing of data and analysis code. Independent research teams attempting to reproduce the same study based on its free-text description alone may produce a range of different implementations that deviate from the original study, and these deviations can have material impact on the size and composition of the study population. Sharing analytical code supported by a common data model allows for the reproduction of a study in an unambiguous way thereby preserving initial logic and study design choices. As reproducibility increasingly becomes an expectation for observational research, standardized open-source tools and practices that enable transparency and consistency should facilitate reproducible research and build trust in the real-world evidence generation process.

AUTHOR CONTRIBUTIONS

AO, GH, MC, PBR, and YA contributed to conception and design of the study. All authors participated in data acquisition. AO and PBR analyzed the data and drafted the manuscript. ASK, AO, AG, AS, DMK, EAV, FL, GH, JPG, JI, LYHL, LZ, MAS, NR, NP, PBR, SLD, and SF interpreted the results. ASK, AO, AG, AS, CL, CK, CB, DMK, DQN, EAV, EM, FL, GH, GR, IR, JPG, JW, JNS, JI, JMB, KS, KEL, LYHL, LZ, MIN, MAS, MC, NR, MP, PS, PBR, SLD, SF, TO, VP, YA, and WAB participated in reviewing the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL

The New England Institutional Review Board determined that studies conducted in CCAE were exempt from study-specific Institutional Review Board review, as these studies do not qualify as human subject research.

SUPPLEMENTARY MATERIAL

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

CONFLICT OF INTEREST STATEMENT

SLD reports grants from Alnylam Pharmaceuticals, Inc, Astellas Pharma, Inc, AstraZeneca Pharmaceuticals LP, Biodesix, Boehringer Ingelheim International GmbH, Celgene Corporation, Eli Lilly and Company, Genentech Inc., Gilead Sciences Inc., GlaxoSmithKline PLC, Innocrin Pharmaceuticals Inc., IQVIA Inc., Janssen Pharmaceuticals, Inc., Kantar Health, MDxHealth, Merck & Co., Inc., Myriad Genetic Laboratories, Inc., Novartis International AG, Parexel International Corporation through the University of Utah or Western Institute for Veteran Research outside the submitted work. AS, CB, DMK, EAV, GR, JI, JNS, JPG, JW, MC, SF, and PBR are employees of Janssen Research and Development LLC and shareholders of Johnson & Johnson (J&J) stock.

DATA AVAILABILITY

The data that support the findings of this study are available to license from IBM at https://www.ibm.com/products/marketscan-research-databases.

REFERENCES

1. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why observational studies should be among the tools used in comparative effectiveness research. *Health Aff (Millwood)* 2010; 29 (10): 1818–25.
2. Holve E, Lopez MH, Scott L, Segal C. A tall order on a tight timeframe: stakeholder perspectives on comparative effectiveness research using electronic clinical data. *J Comp Eff Res* 2012; 1: 441–51.
3. Barba LA, Barba LA, Thiruvathukal GK. Trustworthy computational evidence through transparency and reproducibility. *Comput Sci Eng* 2021; 23 (1): 58–64.
4. Wang S, Verpillat P, Rassen J, Patrick A, Garry E, Bartels D. Transparency and reproducibility of observational cohort studies using large healthcare databases.: transparency and reproducibility in healthcare databases. *Clin Pharmacol Ther* 2016; 99: 325–32.
5. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018; 25 (8): 963–8.
6. Nosek BA, Hardwicke TE, Moshontz H, *et al.* Replicability, robustness, and reproducibility in psychological science. *Annu Rev Psychol* 2022; 73: 719–48.
7. Hardwicke TE, Bohn M, MacDonald K, *et al.* Analytic reproducibility in articles receiving open data badges at the journal *Psychological Science*: an observational study. *R Soc Open Sci* 2021; 8 (1): 201494.
8. Errington TM, Mathur M, Soderberg CK, *et al.* Investigating the replicability of preclinical cancer biology. *ELife* 2021; 10: e71601.

9. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 2015; 349: aac4716.
10. Knottnerus A, Tugwell P. STROBE—a checklist to STrengthen the Reporting of OBservational studies in Epidemiology. *J Clin Epidemiol* 2008; 61 (4): 323.
11. Benchimol EI, Smeeth L, Guttmann A, *et al.* RECORD Working Committee, The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015; 12: e1001885.
12. Kurz X, Perez-Gutthann S; ENCePP Steering Group. Strengthening standards, transparency, and collaboration to support medicine evaluation: ten years of the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). *Pharmacoepidemiol Drug Saf* 2018; 27 (3): 245–52.
13. Wang SV, Pinheiro S, Hua W, *et al.* STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021; 372: m4856.
14. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015; 13: 1.
15. Harris JK, Johnson KJ, Carothers BJ, Combs TB, Luke DA, Wang X. Use of reproducible research practices in public health: a survey of public health analysts. *PLoS One* 2018; 13 (9): e0202447.
16. Hardwicke TE, Wallach JD, Kidwell MC, Bendixen T, Crüwell S, Ioannidis JPA. An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *R Soc Open Sci* 2020; 7 (2): 190806.
17. Albogami Y, Cusi K, Daniels MJ, Wei Y-JJ, Winterstein AG. Glucagon-like peptide 1 receptor agonists and chronic lower respiratory disease exacerbations among patients with type 2 diabetes. *Dia Care* 2021; 44: 1344–52.
18. Wang SV, Schneeweiss S, Berger ML, *et al.*; Joint ISPE-ISPOR Special Task Force on Real World Evidence in Health Care Decision Making. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. *Value Health* 2017; 20 (8): 1009–22.
19. ATLAS. 2022. https://github.com/OHDSI/Atlas. Accessed April 23, 2022.
20. Fletcher S, Islam MZ. Comparing sets of patterns with the Jaccard index. *AJIS* 2018; 22. https://doi.org/10.3127/ajis.v22i0.1538.
21. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput* 2009; 38: 1228–34.
22. Gold S, Lehmann H, Schilling L, Lutters W. Practices, norms, and aspirations regarding the construction, validation, and reuse of code sets in the analysis of real-world data. *Medrxiv* 2021; https://doi.org/10.1101/2021.10.14.21264917.
23. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. *PLoS One* 2018; 13 (2): e0192033.
24. Hornbrook MC, Whitlock EP, Berg CJ, *et al.* Development of an algorithm to identify pregnancy episodes in an integrated health care delivery system. *Health Serv Res* 2007; 42: 908–27.
25. Hardy JR, Holford TR, Hall GC, Bracken MB. Strategies for identifying pregnancies in the automated medical records of the General Practice Research Database. *Pharmacoepidemiol Drug Saf* 2004; 13 (11): 749–59.
26. Devine S, West S, Andrews E, *et al.* The identification of pregnancies within the general practice research database. *Pharmacoepidemiol Drug Saf* 2010; 19: 45–50.
27. Li Q, Andrade SE, Cooper WO, *et al.* Validation of an algorithm to estimate gestational age in electronic health plan databases. *Pharmacoepidemiol Drug Saf* 2013; 22 (5): 524–32.
28. Margulis AV, Setoguchi S, Mittleman MA, Glynn RJ, Dormuth CR, Hernández-Díaz S. Algorithms to estimate the beginning of pregnancy in administrative databases: estimating the beginning of pregnancy. *Pharmacoepidemiol Drug Saf* 2013; 22 (1): 16–24.
29. Mikolajczyk RT, Kraut AA, Garbe E. Evaluation of pregnancy outcome records in the German Pharmacoepidemiological Research Database (GePaRD). *Pharmacoepidemiol Drug Saf* 2013; 22 (8): 873–80.
30. Haibe-Kains B, Adam GA, Hosny A, *et al.*; Massive Analysis Quality Control (MAQC) Society Board of Directors. Transparency and reproducibility in artificial intelligence. *Nature* 2020; 586 (7829): E14–6.
31. Peng RD, Hicks SC. Reproducible research: a retrospective. *Annu Rev Public Health* 2021; 42: 79–93.
32. Gottesman O, Kuivaniemi H, Tromp G, and The eMERGE Network, *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* 2013; 15 (10): 761–71.
33. Goldacre B. All BMJ research papers should share their analytic code. *BMJ* 2016; 352: i886.
34. Goldacre B, Morton CE, DeVito NJ. Why researchers should share their analytic code. *BMJ* 2019; 367: l6365.