Process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs (PRINCIPLED): considerations from the FDA Sentinel Innovation Center

Rishi J Desai,[1] Shirley V Wang,[1] Sushama Kattinakere Sreedhara,[1] Luke Zabotka,[1] Farzin Khosrow-Khavar,[1] Jennifer C Nelson,[2] Xu Shi,[3] Sengwee Toh,[4] Richard Wyss,[1] Elisabetta Patorno,[1] Sarah Dutcher,[5] Jie Li,[5] Hana Lee,[5] Robert Ball,[5] Gerald Dal Pan,[5] Jodi B Segal,[6] Samy Suissa,[7] Kenneth J Rothman,[8] Sander Greenland,[9] Miguel A Hernán,[10] Patrick J Heagerty,[11] Sebastian Schneeweiss[1]

For numbered affiliations see end of the article

Correspondence to: R J Desai rdesai@bwh.harvard.edu (or @RishiDesai11 on Twitter; ORCID 0000-0003-0299-7273)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2024;384:e076460 http://dx.doi.org/10.1136/bmj-2023-076460

Accepted: 11 December 2023

This report proposes a stepwise process covering the range of considerations to systematically consider key choices for study design and data analysis for non-interventional studies with the central objective of fostering generation of reliable and reproducible evidence. These steps include (1) formulating a well defined causal question via specification of the target trial protocol; (2) describing the emulation of each component of the target trial protocol and identifying fit-for-purpose data; (3) assessing expected precision and conducting diagnostic evaluations; (4) developing a plan for robustness assessments including deterministic sensitivity analyses, quantitative bias analyses, and net bias evaluation; and (5) inferential analyses.

SUMMARY POINTS

Non-interventional studies (also referred to as observational studies) conducted using healthcare data that are generated during provision of routine clinical care (including health insurance claims and electronic health records) provide an opportunity to fill in evidence gaps for questions not answered by randomized trials

Despite several assessment and guideline tools available to evaluate the validity of such non-interventional studies, none proposes a practical guide for the conduct and analysis of these studies

PRINCIPLED (process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs) is a stepwise process proposed to systematically consider key choices for study design and data analysis for non-interventional studies

The process outlined here can inform the conduct of non-interventional studies, facilitate transparent communications between various stakeholders, and could motivate similar considerations for the clinical research community

Non-interventional studies, also referred to as observational studies, are conducted using real world data sources typically including healthcare data that are generated during provision of routine clinical care (including health insurance claims and electronic health records). These studies provide an opportunity to fill in evidence gaps for questions that have not been answered by randomized trials.[1] However, generating decision grade evidence from healthcare data requires a robust causal framework to avoid introducing bias. Numerous tools aimed at improving the conduct or reporting of these non-interventional studies are available. Broad guidance documents discuss the methodology for non-interventional studies—such as the best practices for pharmacoepidemiological safety studies by the Food and Drug Administration (FDA)[2] and the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (EncEPP) guide on methodological standards in pharmacoepidemiology.[3] Quality assessment tools such as ROBINS-I[4] and GRACE checklist[5] assist with the evaluation of bias in published studies. Reporting tools such as RECORD-PE[6] and STaRT-RWE[7] provide checklists or structured templates to facilitate transparency in protocol reporting and reproducibility. Finally, the harmonized protocol template HARPER[8] is supported by regulators to improve communication of key study parameters in non-interventional studies, and is deposited with protocol registration websites (eg, the Open Science Foundation's OSF.io and European Medicines Agency's ENcEPP.eu).[9 10] While useful for their specific purposes, these tools are not explicitly intended to guide the design and conduct of non-interventional studies that evaluate drug safety and effectiveness using healthcare data.

Other frameworks such as LEGEND[11] and the causal roadmap[12] outline some broad general principles for evidence generation. However, they provide limited practical guidance on critical aspects of the process of evidence generation, including determining fitness-for-purpose of the data source, registering study protocols, considering principled adaptations over the course of a study, and planning robustness evaluations. To that end, we present a stepwise process covering these key choices with respect to design and analysis that can influence the validity of such studies. We initiate our discussion by considering the FDA Sentinel system, a

national, postmarketing active surveillance system for drug products[13] using large volumes of healthcare data from insurance claims and electronic health records as a representative use case. The five step process outlined in this report covers formulating a well defined causal question via specification of the target trial protocol; describing the emulation of each component of the target trial protocol and identifying fit-for-purpose data source; assessing expected precision and conducting diagnostic evaluations; developing a plan for robustness assessments including deterministic sensitivity analyses, quantitative bias analyses, and net bias evaluation; and inferential analyses.

Overview of the proposed process guide

PRINCIPLED (process guide for inferential studies using healthcare data from routine clinical practice to evaluate causal effects of drugs) is a five step process to help ask and answer a causal question regarding drug treatment effects using healthcare data. We explicitly differentiate between a study planning phase (steps 1-4) where no inference is made, and a study analysis phase (step 5) where inferential analyses are conducted with the intent to draw causal inferences. Figure 1 shows an overview of the proposed steps. Sections below discuss each of the steps in detail. We illustrate the operationalization of each step through an example of the evaluation of sodium-glucose cotransporter-2 (SGLT-2) inhibitors, drugs used for type 2 diabetes treatment, with respect to the known safety concern of genital infections.[14] While this process considers an iterative general approach to resolve issues as they arise during conduct of non-interventional studies, specific situations could necessitate deliberate deviation from these steps. Even in situations where the process cannot be fully implemented, a reasonable study could still be conducted, but certain trade-offs might need to be made.

Step 1: Formulate a causal question via specification of the target trial protocol

Asking the right question in the right manner constitutes the first step in any process for causal inference about treatment effects from observed data.[15 16] A practical way to ask a causal question in non-interventional studies is to specify a protocol of the target trial—the pragmatic trial that would answer the causal question.[17 18] Among the key elements of the target trial protocol that need to be specified are eligibility criteria, treatment strategies, primary outcome(s) of interest, treatment assignment, start and end of the follow-up, and causal contrast (eg, intention-to-treat or per protocol effect). Precise specification of the target trial protocol is critical because it has direct implications in analysis and interpretation. For instance, specified eligibility criteria determine the population to which the results would apply. Table 1 summarizes the basic target trial protocol for our case example study.

Step 2: Describe the emulation of each component of the target trial protocol and identify a fit-for-purpose data source

Specifying the key components of the target trial protocol in step 1 clarifies a list of the data elements necessary to emulate it. Next, confounders that are necessary to emulate baseline randomization should be identified. Causal diagrams, such as causal directed acyclic graphs, are useful to make decisions

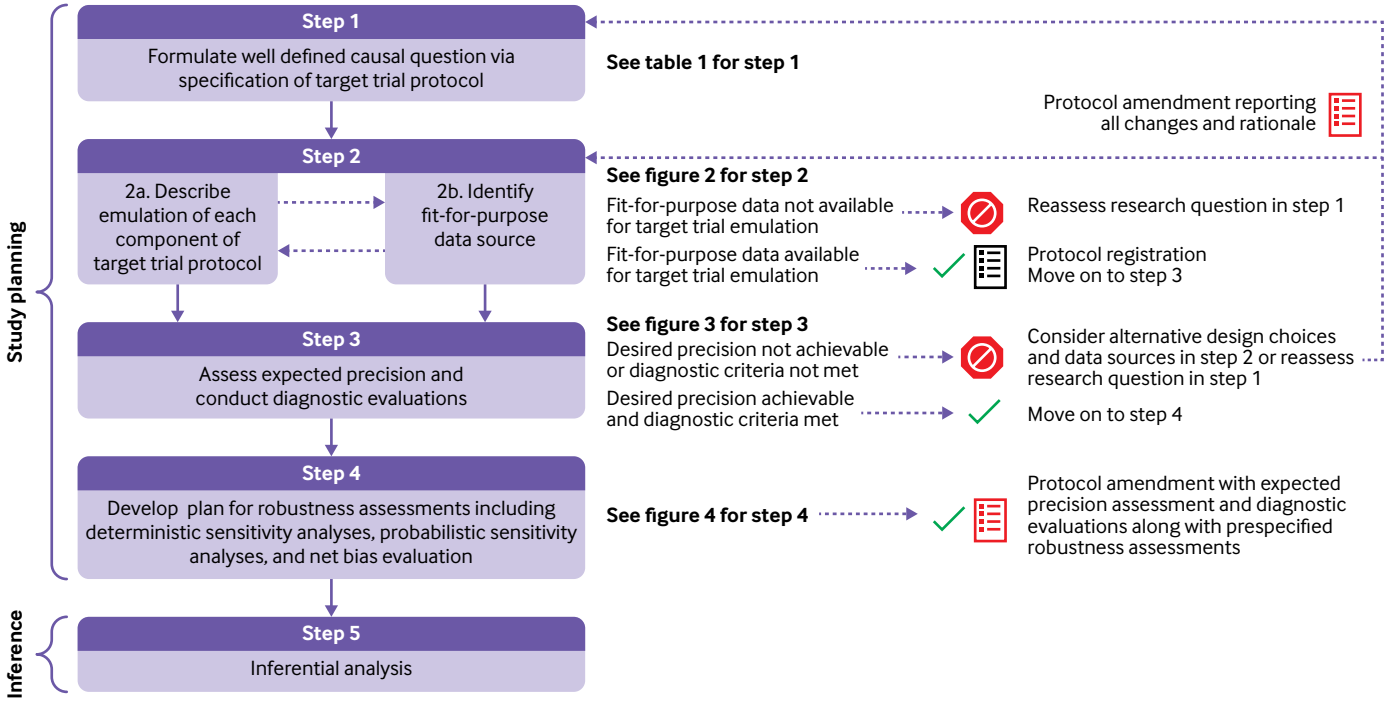

Fig 1 | Overview of the process guide for inferential studies using healthcare data from routine clinical practice

Table 1 | Target trial protocol for case example study evaluating the effect of sodium-glucose cotransporter-2 (SGLT-2) inhibitors on genital infections

Element	Specification	Emulation using real world data sources
Eligibility criteria	Patients with type 2 diabetes mellitus; aged ≥65 years; no use of study drug treatments before randomization; no history of end stage renal disease, HIV, or genital infections; continuous Medicare A, B, D enrolment for six months and recorded glycated hemoglobin (HbA_{1c}) test results in electronic health records in six months before treatment initiation	Same as target trial
Treatment strategies	Initiation of (1) SGLT-2 inhibitors (canagliflozin, dapagliflozin, empagliflozin); or (2) DPP-4 inhibitors (alogliptin, linagliptin, saxagliptin, sitagliptin). Under both strategies, use of antidiabetic treatment after initiation is left to physician and patients' discretion	Same as target trial
Treatment assignment	Randomized, non-blinded	Non-blinded and assumed to be randomized within levels of measured confounders*
Follow-up start (time 0)	At assignment	Same as target trial
Follow-up end	First of administrative end of follow-up (day 365), loss to follow-up, death, or outcome occurrence	Same as target trial
Primary outcome	Genital infections	Same as target trial
Causal contrast	Intention-to-treat effect (effect of being assigned to the treatment)	Observational analogue of intention-to-treat effect

SGLT-2=sodium-glucose cotransporter-2; DPP-4=dipeptidyl peptidase-4; HbA1c=glycated hemoglobin.
*Measured confounders include demographics (age, sex, race, socioeconomic status markers), diabetes severity related variables including microvascular and macrovascular complications, measures related to diabetes control such as HbA_{1c}, comorbid conditions, cotreatments, markers for healthy behavior, and healthcare use.

about confounder selection when sufficient content knowledge is available.[19 20] Importantly, adjustment for colliders and instrumental variables should be avoided.[21]

Once all data elements are outlined, investigators need to describe the emulation of each component of the target trial protocol by providing a precise description of variable definitions, including all codes and algorithms used for eligibility criteria, treatment strategies (including treatment initiation and discontinuation), outcomes, and confounders (step 2a). Data analyses that would be implemented if the data from the target trial were available should also be described in detail. Structured protocol templates such as STaRT-RWE[7] and HARPER[8] are available to assist with transparent reporting of the study protocol. A design diagram is suggested to summarize visually the longitudinal design aspects of a study.[22]

Next, investigators need to identify fit-for-purpose data sources that contain all data elements needed for successful emulation of the target trial (step 2b). Target trial specification is an iterative process that depends on the availability of data to support the emulation. If certain data elements are not included in the data source being considered, investigators can consider alternate data sources.

As an example of selection of fit-for-purpose data, we consider the Sentinel system, which contains structured data from health insurance claims representing 844 million person years of observation between 2000 and 2021 across a large network of data providers,[23] and is increasingly being enriched with insurance claims and linked data from electronic health records.[24] Figure 2 outlines an approach to assess the fitness of purpose that is compatible with FDA draft guidance to industry on real world data.[25] Two key considerations are data relevance and data reliability. For determination of relevance, we consider the context of Sentinel where most of the data come from insurance claims, and ancillary sources (including electronic health records) provide opportunities for augmentation. In this case, relevance determination depends on a series of questions focused on measurement characteristics of four variable types central to the research question of interest in insurance claims data: eligibility criteria, outcome, treatment, and key confounders. If measurement of any of these variables is deemed to be insufficient, augmentation of insurance claims with alternate sources such as linked electronic health records would be needed. We describe below the specific nuances when considering these four key questions.

- Question 1: Can the eligibility criteria be emulated with sufficient accuracy?

  Certain eligibility criteria specified in the target trial protocol (eg, some medical conditions) might not be explicitly identifiable in insurance claims and a previously validated phenotyping algorithm might not be available. In these circumstances, linkage to electronic health records will be needed for development and validation of phenotyping algorithms identifying the health conditions of interest using claims based proxy information.

  For instance, heart failure subtypes of preserved and reduced ejection fraction are not directly identifiable in insurance claims owing to lack of ejection fraction measurements. A probabilistic phenotyping algorithm based on Medicare claims for identifying ejection fraction subtypes for heart failure was developed using Medicare claims linked to electronic health records from the Mass General Brigham healthcare system. It demonstrated overall accuracy of 83% in differentiating between preserved and reduced ejection fraction subtypes.[26] This model facilitated deployment of this algorithm in national Medicare claims data to study drug treatment outcomes for these specific populations of interest.[27 28] In circumstances where a developed algorithm demonstrates suboptimal performance, limiting

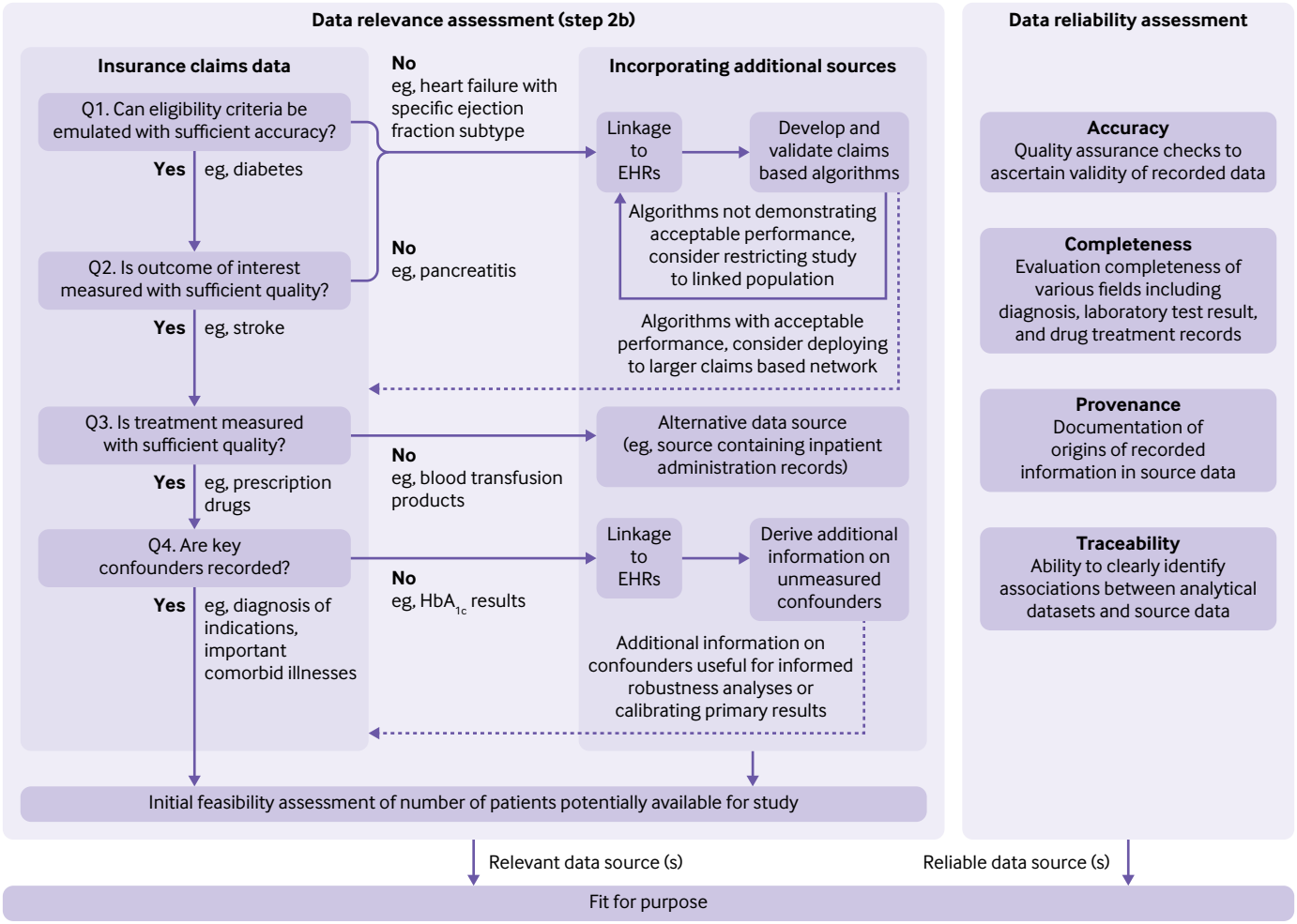

Fig 2 | Determining fit-for-purpose data sources (step 2b of the process guide for inferential studies using healthcare data from routine clinical practice). HbA_{1c}=glycated hemoglobin; EHR=electronic health records. *Quality=accuracy with respect to timing and completeness for treatments; positive predicted value, sensitivity, specificity for binary outcomes; proportion missing for continuous outcomes; accurate onset for time to event outcomes; and availability of long term follow-up data for latent outcomes

the analysis to individuals with linked data from insurance claims and electronic health records available and a pre-treatment measurement of the eligibility criteria might be needed to prevent bias at the expense of transportability.

- Question 2: Is the outcome of interest measured with sufficient quality?

  The quality of outcome measurement depends on positive predicted value for binary outcomes, proportion missing for continuous outcomes, and accurate onset for time-to-event outcomes. Typically, serious medical conditions (eg, stroke) might be adequately recorded in insurance claims[29]; but other outcomes are not, including those that require confirmatory laboratory test results (eg, acute pancreatitis[30]) or contextual information from free text notes (eg, suicidal ideation[31]). For such outcomes, data augmentation through linkage of insurance claims with electronic health records is required.

Outcome-identifying algorithms (including those using only claims based information) can be developed, improved, and validated based on chart reviews using linked electronic health records. If an algorithm using only claims based information shows acceptable performance, such an algorithm can be applied to the larger insurance claims data source. In cases where claims based algorithms are insufficient but electronic health record sources provide sufficient augmentation to identify the outcome, researchers could consider restricting their population to patients with claims-electronic health records linked records. Judgments on the quality required for an algorithm to be considered sufficient for use in inference can be subjective; however, implementing a simplified rule on performance parameters (eg, ≤85% positive predicted value) might not be helpful. Whether to proceed with the analysis is a multifaceted decision and considers factors such as the urgency of information needed

and the severity of the adverse event. Knowing the measurement characteristics through validation in linked electronic health records, even when they are suboptimal, will enable quantitative bias analysis.[32] More details on quantitative bias analysis are given below in step 4. In analyses that go across a network of databases, the transportability of measurement algorithms and the measurement qualities across databases might need to be demonstrated.

- Question 3: Is the treatment measured with sufficient quality?

  Quality of measurement for a particular treatment refers to the accuracy of recording in insurance claims data with respect to the timing and completeness. For many products such as outpatient prescription drug treatments, insurance claims are generally sufficient to capture treatment through outpatient pharmacy dispensing records. However, an example treatment that is often insufficiently recorded in claims is blood transfusion products.[33] In such circumstances, alternate data sources that have information on inpatient administrations are needed to answer the research question. If dynamic treatment strategies are being compared, the time-varying clinical factors used to define the strategies over time should also be available.[34]

- Question 4: Are key confounders recorded?

  If a strong confounder is not adequately measured in insurance claims, data augmentation with electronic health records or laboratory test results might need to be considered. For example, baseline glycated hemoglobin (HbA_{1c}) test results for a study comparing two glucose-lowering drug treatments with respect to an adverse outcome might require augmentation. Added information on confounders achieved through augmentation might be useful to assess the potential for uncontrolled confounding,[35] and for performing additional analyses such as statistical calibration of the study results to incorporate knowledge about unmeasured confounders.[36]

Data sources meet the basic criteria for relevance, potentially through various augmentation strategies if needed, when they provide explicitly characterized eligibility criteria, primary outcomes, treatment, and key confounders. Additionally, initial feasibility assessment of the number of patients potentially available for the study might be needed to ensure relevance. For example, such assessments could include an initial evaluation of the number of new users of study drug treatments of interest in the data source(s) being considered.

The second aspect for fitness-for-purpose of a data source is data reliability, which includes assessments of accuracy, completeness, provenance, and traceability of the source data (fig 2).[25] Within Sentinel, these evaluations are performed upstream when converting raw data from contributing sources to the Sentinel common data model—which is then used for all subsequent analyses.[37] Data sources that meet both relevance and reliability criteria can be considered fit for purpose for the study question of interest.

If emulation of each component of the target trial protocol is not feasible with the data source being considered, investigators can reassess the question in step 1 by specifying a modified target trial protocol that requires a different set of data elements while still asking a causal question of interest. Investigators are encouraged to record all assessments of data relevance and data reliability to trace key design decisions leading to selection of fit-for-purpose data that can support the corresponding trial emulation.

If emulation of each component of the target trial protocol is feasible with the data source being considered, investigators should consider registration of the study protocol at this stage before proceeding with assessment of expected precision and diagnostic evaluations (step 3). An alternative to protocol registration is publication of the target trial protocol along with the annotated computer code while making the data available to interested investigators whenever feasible. Pre-registration of protocols and data sharing agreements can serve as deterrent to data dredging, which is a common concern with analyses of healthcare data.[38]

For the case example study, demographics (age, sex, race, socioeconomic status markers); variables related to diabetes severity including microvascular and macrovascular complications; measures related to diabetes control such as HbA_{1c}, comorbid conditions, co-treatments, markers for healthy behavior, and healthcare use were considered confounders owing to their likely association with treatment choice and outcome risk. We describe the emulation of each component of the target trial protocol by providing a precise description of the operationalization of variable definitions, including all codes and algorithms, using the HARPER[8] template (web appendix 2). For statistical analysis, we estimated the hazard ratio (averaged over the follow-up period) via a Cox model adjusted for baseline confounding with propensity score stratification and weighting,[39 40] as in previous studies with low incidence of treatment initiation and rare safety outcomes.[41] Other adjustment methods, such as parametric g formula or inverse probability weighting, might be required when emulating trials with sustained treatment strategies and thus with time-varying treatments.[42] We also specified analyses in groups stratified by sex, age, and baseline risk factors for infections as subgroup analyses of interest to evaluate potential effect measure modification by these characteristics.

Appendix figure 1 answers questions 1-4 to provide clarity on likely fit-for-purpose data for our case example. Briefly, outcome and treatment are well captured in Medicare claims; however, linkage to electronic health records could be important to ascertain clinical factors that are used as eligibility

criteria or confounders. In this case example, we used US Medicare Fee For Service claims data from parts A, B, D that are deterministically linked by health insurance claim numbers, date of birth, and sex (linkage success rate 99.2%) to electronic health records from the Mass General Brigham healthcare system in Boston.

Step 3: Assess expected precision and conduct diagnostic evaluations

After clearly specifying all design choices and registering a study protocol, the next important design component is assembling the study population using all eligibility criteria to assess expected precision and to conduct diagnostic evaluations. These evaluations could allow for principled study adaptations, yet little formal guidance exists regarding this activity. We fill this gap by outlining a systematic approach in figure 3.

- Step 3a: Assess expected precision

  For emerging safety signals where effect size is likely not known, the decision to proceed with analyses should depend on the importance of the information gained from a public health perspective.[43] However, during the planning phase, it might be helpful to gauge the expected precision based on the selected data source and design choices to determine if adjustments are needed to achieve desired level of precision.[44] Based on the outcome counts and sizes of two treatment groups, researchers can estimate the variance of the log risk ratio using well known formulas and assumptions about the magnitude of the risk ratio.[44] We provide an R function to estimate expected precision based on sizes of two treatment groups and combined outcome counts across two groups as supplemental material (web appendix 3).

- Step 3b: Diagnostic evaluations

  Diagnostic evaluations are key components of non-interventional studies because they can alert researchers to potential violations of the core assumptions of causal inference. For instance, examining distribution of baseline characteristics in treatment groups being compared is an important diagnostic to detect positivity violations.[45] Evaluating average length of time during which patients adhere to their assigned treatment strategies and examining characteristics of patients who deviate from the treatment strategies could alert researchers to the possibility of informative censoring, which could threaten exchangeability. Other analysis specific diagnostic criteria might also be helpful. For instance, when using analyses based on propensity scores, evaluating baseline covariate balance after

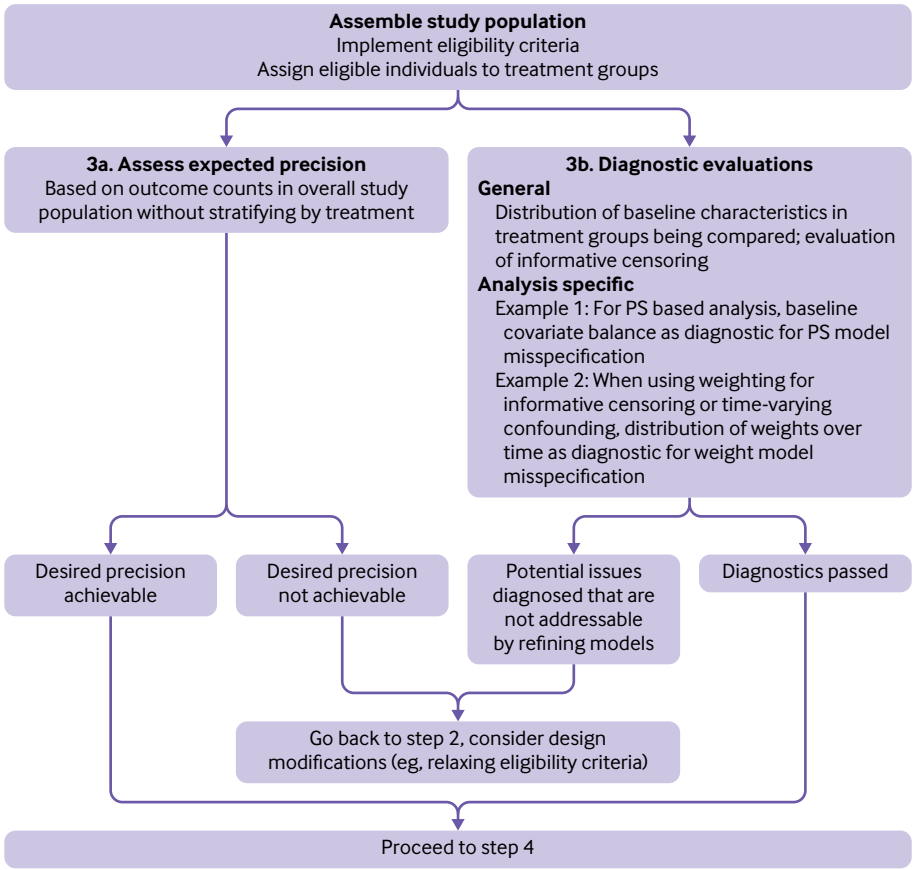

Fig 3 | Assessing expected precision and conducting diagnostic evaluations (step 3 of the process guide for inferential studies using healthcare data from routine clinical practice). PS=propensity score

conditioning on the propensity score could serve as a diagnostic for model misspecification.[40 46 47] If inverse probability weighting is used to adjust for informative censoring or time-varying confounding, evaluating distribution of weights over time could serve as a diagnostic for weight model misspecification.[48] For analysis specific diagnostics, refining modelling choices could lead to resolution of issues.

If the assessment indicates lower than desirable precision or diagnostic evaluations indicate violations of core causal inference assumptions that cannot be resolved by refining modelling choices, investigators can consider going back to step 2 and changing some design choices, such as eligibility criteria or choice of the comparator group, before proceeding. This suggestion is analogous to an amendment of the study protocol that is common in prospective randomized trials in response to extraneous factors such as recruiting challenges.[49] Similar to the guidance regarding protocol amendments for prospective trials, reasons for changes in the protocol of non-interventional studies using secondary healthcare data need to be clearly documented, as well as any changes in the causal contrasts that result from protocol changes. To maintain analyst blinding with respect to the treatment and outcome association and study integrity, researchers should also ensure that protocol amendments are not introduced in response to inferential analysis (step 5).

For our case example in step 3a, the expected 95% confidence interval under an assumed null effect on the relative scale (1.0) of SGLT-2 inhibitors on the risk of genital infections was 0.35 to 1.65. This result is imprecise because only 1498 patients with only 40 outcomes were eligible for analysis. Because the low sample size is partly due to the inclusion criterion of HbA_{1c} test results before initiation of drug treatment (appendix fig 2), we could go back to step 2 and consider relaxing this inclusion criterion, which would increase the number of eligible individuals to 9339 (293 events) with a 95% confidence interval of 0.73 to 1.27. However, relaxing this criterion makes the assumption that not adjusting for HbA_{1c} in the main analysis does not introduce major confounding bias. Appendix table 1 provides a revised target trial table highlighting the one protocol change prompted by assessment of expected precision.

For step 3b, we used this cohort of 9339 patients meeting eligibility criteria per the amended protocol. We estimated the probability of initiating SGLT-2 inhibitors versus DPP-4 (dipeptidyl peptidase-4) inhibitors given baseline patient characteristics (ie, the propensity score) using multivariable logistic regression models, created 50 stratums based on the distribution of propensity scores in patients receiving SGLT-2 inhibitor treatment, and weighted DPP-4 inhibitor initiators proportional to the distribution of SGLT-2 inhibitor initiators in the propensity score stratum into which they fell.[39] As diagnostics for propensity score models, we evaluated distributional overlap (appendix fig 3), weight distribution (appendix fig 4), and covariate balance using standardized differences post-weighting (appendix tables 2 and 3).[40 50] SAS macros used to conduct the analysis and generate diagnostic figures are publicly available.[51] All SAS codes are also posted on https://dev.sentinelsystem.org/projects/IC/repos/ic_ci2_principled/browse.

Step 4: Develop a plan for robustness assessments including deterministic sensitivity analyses, probabilistic sensitivity analyses, and net bias evaluation

Robustness assessments deal with the consistency of evidence with respect to alternative investigator decisions related to study design, measurement, or analysis. As the fourth and final step of study planning, we propose prespecification of robustness assessments. After assessing precision and diagnostic evaluations, investigators probably have additional understanding of the potential threats to the study and can make informed judgments related to the need for specific robustness evaluations. Such prespecified assessments are most useful if they have a clear rationale regarding the specific types of bias they address. Robustness assessments can be broadly categorized into three types, which are detailed below (fig 4).

- Step 4a: Deterministic sensitivity analyses

  Deterministic sensitivity analyses, also known as deterministic quantitative bias analysis, can be viewed as variations of the target trial protocol, where investigators focus on specific design or analytical assumptions and vary them individually to gauge the impact of specific assumptions or design choices on study results. Deterministic sensitivity analysis could focus on highly specific design or measurement choices, such as varying the outcome definition to increase the specificity and evaluate the possibility of bias due to outcome misclassification. They could also involve prespecification of alternate statistical analysis methods.

- Step 4b: Probabilistic sensitivity analyses

  Probabilistic sensitivity analyses, also known as probabilistic quantitative bias analysis, use various probabilistic and simulation approaches to evaluate the impact of various hidden biases on study results, including exposure/outcome misclassification, unmeasured confounders, and selection bias.[35 52] Monte Carlo simulations evaluating potential bias require realistic ranges for bias parameters, for instance, sensitivity and specificity of an outcome identifying algorithm based on existing information such as validation studies.[53] In those simulations, study results are recalculated for each run and then tabulated to provide empirical estimates of expected variation due to uncertainties in exposure or

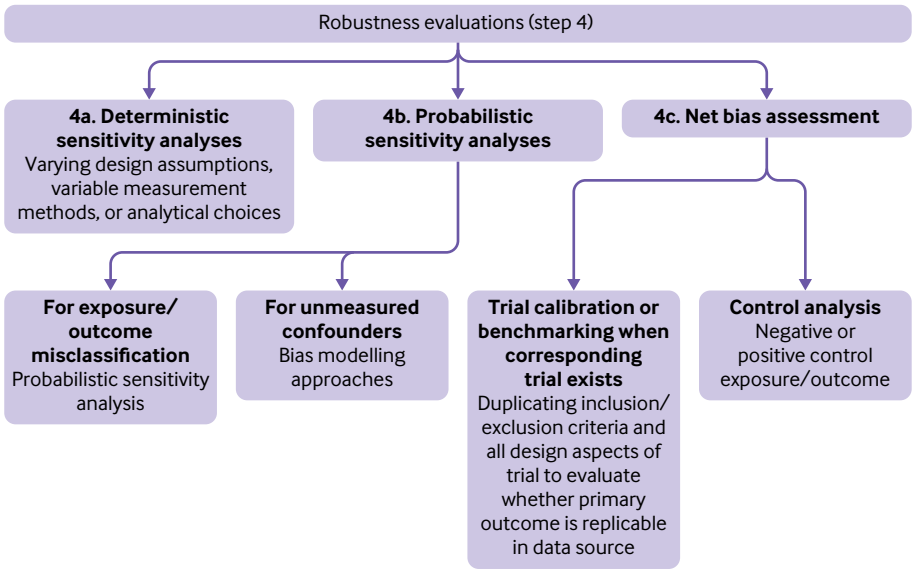

Fig 4 | Robustness evaluations (step 4 of the process guide for inferential studies using healthcare data from routine clinical practice)

outcome identification.[32] Similar bias modelling approaches are available to evaluate the impact of unmeasured confounders on study results based on the strength of association between the exposure and the suspected confounder as well as the outcome and the suspected confounder.[35]

- Step 4c: Net bias assessment

  We use the term "net bias assessment" to describe the approaches that allow investigators to detect presence of bias from multiple sources such as uncontrolled confounding, selection bias, and measurement error. We describe two major types of such assessments.

  Firstly, where possible, investigators should a priori identify and include control outcomes or control exposures that are known to have no associations (negative controls) or well established associations (positive controls) with either the exposure or outcome of interest. Ideally, these control variables will have confounding structure or mechanism of measurement error similar to the effect targeted for study.[54 55] Inability to replicate the known effect sizes in these analyses could alert investigators to the presence of bias.

  Secondly, when a well conducted randomized trial exists for the comparison under investigation with a different primary endpoint or conducted within a more restrictive population, benchmarking or trial calibration might be pursued.[56 57] If investigators are able to replicate results for the primary outcome of such a trial in their data source by using identical inclusion and exclusion criteria and other design elements, it could increase confidence in results under a modified target trial protocol.

We recommend that investigators add expected precision assessment and diagnostic evaluations along with prespecified robustness assessments as amendments to the registered protocol before moving on to step 5. If assessment of expected precision and diagnostic evaluations, which explicitly do not allow any inferential analyses, lead to any meaningful adaptations in the design or measurement, all such changes should also be documented as amendments to the registered protocol before starting the inferential analyses.

For our case example, we specified a deterministic sensitivity analysis (step 4a) to evaluate the impact of outcome misclassification. We defined the outcome after excluding non-specific codes of balanitis and balanoposthitis in male patients and vaginitis and vulvovaginitis in female patients and focusing solely on candida of urogenital sites.

We also specified a quantitative bias analysis (step 4b). To explore the impact of our assumption that HbA_{1c} is not an important confounder, we used HbA_{1c} data in a subset of patients to inform this analysis.[58] Information regarding the distribution of HbA_{1c} in our linked subset and the association between the unmeasured confounder (HbA_{1c}) and outcome (infections) based on prior epidemiological research[59] were used as inputs to calculate adjusted estimates over a range of bias parameters.

Finally, we specified a net bias analysis (step 4c), by assessing hospital admission for heart failure as a positive control outcome. SGLT-2 inhibitors have an established association with a reduced risk of hospital admission for heart failure. This association has been observed consistently across randomized controlled trials including CANVAS, CREDENCE, DAPA-HF, DECLARE-TIMI-58, EMPAREG OUTCOME, EMPEROR-REDUCED, and VERTIS-CV.[60 61] If the set of controlled covariates is sufficient to control confounding (without introducing bias) for both of the outcomes (genital infection and hospital admission for heart failure), a

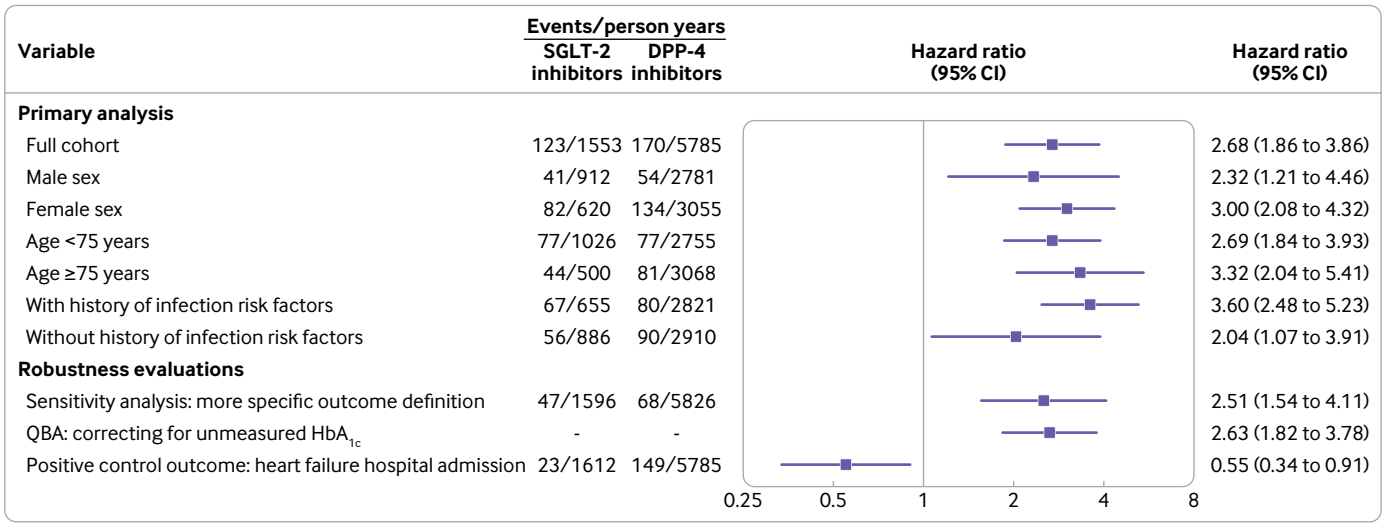

Fig 5 | Results from the primary analysis, subgroup analyses, and robustness evaluations for the case example study evaluating the effect of sodium-glucose cotransporter-2 (SGLT-2) inhibitors on genital infections. The quantitative bias analysis (QBA) presents adjusted results at the values of bias parameters observed in ancillary data (14% uncontrolled hyperglycemia as defined by glycated hemoglobin (HbA_{1c}) >9% in reference group and odds ratio of 1.3 for receipt of SGLT-2 inhibitor treatment). Appendix figure 5 provides results from this quantitative bias analysis over various combinations of bias parameters

finding of robust adjusted association between the exposure and known positive control outcome can provide some reassurance in the observed findings for the genital infection outcome.

Step 5: Inferential analysis

At the end of step 4, all key design elements, measurements, and data analysis plan are prespecified, and inferential data analysis can proceed. The central idea behind structuring the steps in this sequence with a clear demarcation between planning and inference is to avoid design or analysis changes prompted by study results. At the conclusion of inferential analysis and all prespecified robustness evaluations, investigators are well positioned to make sound inferences about the association under investigation.

For our case example study, results are presented in figure 5, which showed a consistently elevated risk of genital infections after initiating SGLT-2 inhibitors versus DPP-4 inhibitors in patients with diabetes across all subgroups and all robustness evaluations. Appendix figure 5 summarizes the quantitative bias analysis for uncontrolled confounding by HbA_{1c} over a range of bias parameters, which indicated that the risk of genital infections with SGLT-2 inhibitors remained elevated even in extreme scenarios of uncontrolled confounding. In net bias analysis, we observed a robust reduction in the risk of the positive control outcome (hospital admission for heart failure), which was expected. Overall, results indicating potentially a greater risk of genital infections with SGLT-2 inhibitors are in line with prior observations from trials and observational studies. In a large meta-analysis of eight phase 3 randomized trials, the pooled relative risk for genital infections was reported to be 3.75 (95% confidence interval 3.00 to 4.67).[62] A previous analysis of US commercial insurance claims reported about a threefold increased risk of genital infections with SGLT-2 inhibitors versus DPP-4 inhibitors.[63]

Conclusion

This report introduces a stepwise process that systematically considers key decision nodes for evaluating causal effects of treatments using healthcare data. The process outlined in this framework can facilitate transparent communications between various stakeholders and motivate critical considerations for the clinical research community.

AUTHOR AFFILIATIONS

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, USA
[2]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA
[3]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[4]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA
[5]US Food and Drug Administration, Silver Spring, MD, USA
[6]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[7]Departments of Epidemiology and Biostatistics, and Medicine, McGill University, Montreal, QC, Canada
[8]Boston University School of Public Health, Boston, MA, USA
[9]Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA, USA
[10]CAUSALab and Departments of Epidemiology and Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA
[11]Department of Biostatistics, University of Washington, Seattle, WA, USA

Contributors: RJD, SVW, ST, JCN, SS, SD, RB, and GDP have leadership roles in the FDA's Sentinel initiative, which is the national active postmarketing surveillance system for medical products in the US. All other authors are invited experts from academia or FDA with many years of combined experience in development of methods informing conduct of non-interventional studies. Coauthors from the

US Food and Drug Administration (FDA) participated in the results interpretation and in the preparation and decision to submit the manuscript for publication.The authors were brought together as a workgroup supported by the FDA Sentinel Innovation Center. The workgroup held 12 teleconference calls between June 2021 and December 2022, which were attended by authors (RJD, SVW, SKS, LZ, FK-K, JCN, XS, ST, RW, EP, SD, JL, HL, RB, GDP, JBS, SS, KJR, SG, MAH, PJH, and SS) to discuss the process and reach a consensus. RJD, SKS, LZ, and FKK conducted the data analysis for the case example study. RJD is the guarantor of the content of this article. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This project was supported by Master Agreement 75F40119D10037 from the FDA. The FDA approved the study protocol used in the illustrative example shown in web appendix 2, including statistical analysis plan and reviewed and approved this manuscript. The FDA had no role in data collection, management, or analysis. The views expressed are those of the authors and not necessarily those of the FDA.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the FDA for the submitted work. RJD reports serving as principal investigator on investigator initiated grants to the Brigham and Women's Hospital from Novartis, Vertex, and Bayer on unrelated projects. SS is co-principal investigator of an investigator initiated grant to the Brigham and Women's Hospital from Boehringer Ingelheim unrelated to the topic of this study, and is a consultant to Aetion, a software manufacturer of which he owns equity; his interests were declared, reviewed, and approved by the Brigham and Women's Hospital and Mass General Brigham HealthCare System in accordance with their institutional compliance policies. RB is an author on US Patent 9 075 796 (on text mining for large medical text datasets and corresponding medical text classification using informative feature selection), which at present is not licensed and does not generate royalties. JCN reports research funding from Moderna for service on their safety monitoring committee.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Concato J, Corrigan-Curay J. Real-World Evidence - Where Are We Now? *N Engl J Med* 2022;386:1680-2. doi:10.1056/NEJMp2200089
2 Food and Drug Administration. Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets. 2013. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/best-practices-conducting-and-reporting-pharmacoepidemiologic-safety-studies-using-electronic.
3 European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. ENCePP Guide on Methodological Standards in Pharmacoepidemiology. 2022 https://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml.
4 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919. doi:10.1136/bmj.i4919
5 Dreyer NA, Bryant A, Velentgas P. The GRACE checklist: a validated assessment tool for high quality observational studies of comparative effectiveness. *J Manag Care Spec Pharm* 2016;22:1107-13. doi:10.18553/jmcp.2016.22.10.1107
6 Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;363:k3532. doi:10.1136/bmj.k3532
7 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. doi:10.1136/bmj.m4856
8 Wang SV, Pottegård A, Crown W, et al. HARmonized Protocol Template to Enhance Reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. *Value Health* 2022;25:1663-72. doi:10.1016/j.jval.2022.09.001
9 Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf* 2017;26:1033-9. doi:10.1002/pds.4297
10 Orsini LS, Berger M, Crown W, et al. Improving Transparency to Build Trust in Real-World Secondary Data Studies for Hypothesis Testing-Why, What, and How: Recommendations and a Road Map from the Real-World Evidence Transparency Initiative. *Value Health* 2020;23:1128-36. doi:10.1016/j.jval.2020.04.002
11 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 Assoc* 2020;27:1331-7. doi:10.1093/jamia/ocaa103
12 Dang LE, Gruber S, Lee H, et al. A causal roadmap for generating high-quality real-world evidence. *J Clin Transl Sci* 2023;7:e212. doi:10.1017/cts.2023.635
13 Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. *Clin Pharmacol Ther* 2016;99:265-8. doi:10.1002/cpt.320
14 Food and Drug Administration. FDA warns about rare occurrences of a serious infection of the genital area with SGLT2 inhibitors for diabetes. 2018. https://www.fda.gov/drugs/drug-safety-and-availability/fda-warns-about-rare-occurrences-serious-infection-genital-area-sglt2-inhibitors-diabetes.
15 Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am J Public Health* 2018;108:616-9. doi:10.2105/AJPH.2018.304337
16 E9(R1) Statistical Principles for Clinical Trials. Addendum: Estimands and Sensitivity Analysis in Clinical Trials. 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical
17 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. doi:10.1093/aje/kwv254
18 Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Emulating a target trial in case-control designs: an application to statins and colorectal cancer. *Int J Epidemiol* 2020;49:1637-46. doi:10.1093/ije/dyaa144
19 Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol* 2002;155:176-84. doi:10.1093/aje/155.2.176
20 Tennant PWG, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol* 2021;50:620-32. doi:10.1093/ije/dyaa213
21 VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol* 2019;34:211-9. doi:10.1007/s10654-019-00494-6
22 Schneeweiss S, Rassen JA, Brown JS, et al. Graphical depiction of longitudinal study designs in health care databases. *Ann Intern Med* 2019;170:398-406. doi:10.7326/M18-3079
23 Food and Drug Administration. Sentinel Initiative. 2022. https://www.sentinelinitiative.org/about/key-database-statistics
24 Desai RJ, Matheny ME, Johnson K, et al. Broadening the reach of the FDA Sentinel system: A roadmap for integrating electronic health record data in a causal analysis framework. *NPJ Digit Med* 2021;4:170. doi:10.1038/s41746-021-00542-0
25 Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry. 2021. fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory.
26 Desai RJ, Lin KJ, Patorno E, et al. Development and preliminary validation of a Medicare claims–based model to predict left ventricular ejection fraction class in patients with heart failure. *Circ Cardiovasc Qual Outcomes* 2018;11:e004700. doi:10.1161/CIRCOUTCOMES.118.004700
27 Bhatt AS, Vaduganathan M, Zhuo M, Fu EL, Solomon SD, Desai RJ. Risk of Acute Kidney Injury Among Older Adults With Heart Failure and With Reduced Ejection Fraction Treated With Angiotensin-Neprilysin Inhibitor vs Renin-Angiotensin System Inhibitor in Routine Clinical Care. *J Card Fail* 2023;29:138-46. doi:10.1016/j.cardfail.2022.09.004
28 Desai RJ, Solomon SD, Vaduganathan M. Rates of Spironolactone Initiation and Subsequent Hyperkalemia Hospitalizations in Patients with Heart Failure with Preserved Ejection Fraction Following the TOPCAT trial: A Cohort Study of Medicare Beneficiaries. *J Card Fail* 2022;28:1035-9. doi:10.1016/j.cardfail.2022.01.012
29 Wahl PM, Rodgers K, Schneeweiss S, et al. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. *Pharmacoepidemiol Drug Saf* 2010;19:596-603. doi:10.1002/pds.1924
30 Floyd JS, Bann MA, Felcher AH, et al. Validation of acute pancreatitis among adults in an integrated healthcare system. *Epidemiology* 2023;34:33-7. doi:10.1097/EDE.0000000000001541
31 Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci Rep* 2018;8:7426. doi:10.1038/s41598-018-25773-2
32 Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969-85. doi:10.1093/ije/dyu149

33 Sentinel Initiative. Expansion of the US FDA Sentinel System to inpatient blood transfusion data from Hospital Corporation of America: new surveillance options. 2017. https://www.sentinelinitiative.org/sites/default/files/Sentinel-ICPE-2017-Presentation-HCA-Data-Exploration.pdf
34 Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *Int J Biostat* 2010;6:18. doi:10.2202/1557-4679.1212
35 Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006;15:291-303. doi:10.1002/pds.1200
36 Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;162:279-89. doi:10.1093/aje/kwi192
37 Sentinel Operations Center. Sentinel Data Quality Assurance Practices. 2017. https://www.sentinelinitiative.org/sites/default/files/data/distributed-database/Sentinel_DataQAPractices_Memo.pdf
38 Smith GD, Ebrahim S. Data dredging, bias, or confounding, They can all get you into the BMJ and the Friday papers. *BMJ* 2002;325:1437-8. doi:10.1136/bmj.325.7378.1437
39 Desai RJ, Rothman KJ, Bateman BT, Hernandez-Diaz S, Huybrechts KF. A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology* 2017;28:249-57. doi:10.1097/EDE.0000000000000595
40 Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* 2019;367:l5657. doi:10.1136/bmj.l5657
41 Patorno E, Huybrechts KF, Bateman BT, et al. Lithium use in pregnancy and the risk of cardiac malformations. *N Engl J Med* 2017;376:2245-54. doi:10.1056/NEJMoa1612222
42 Smith LH, García-Albéniz X, Chan JM, et al. Emulation of a target trial with sustained treatment strategies: an application to prostate cancer using both inverse probability weighting and the g-formula. *Eur J Epidemiol* 2022;37:1205-13. doi:10.1007/s10654-022-00929-7
43 Hernán MA. Causal analyses of existing databases: no power calculations required. *J Clin Epidemiol* 2022;144:203-5. doi:10.1016/j.jclinepi.2021.08.028
44 Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology* 2018;29:599-603. doi:10.1097/EDE.0000000000000876
45 Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. *Pharmacoepidemiol Drug Saf* 2021;30:1471-85. doi:10.1002/pds.5338
46 Webster-Clark M, Stürmer T, Wang T, et al. Using propensity scores to estimate effects of treatment initiation decisions: State of the science. *Stat Med* 2021;40:1718-35. doi:10.1002/sim.8866
47 Wyss R, Ellis AR, Brookhart MA, et al. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *Am J Epidemiol* 2014;180:645-55. doi:10.1093/aje/kwu181
48 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656-64. doi:10.1093/aje/kwn164
49 Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *J Pharmacol Pharmacother* 2010;1:100-7. doi:10.4103/0976-500X.72352
50 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;38:1228-34doi:10.1080/03610910902859574.
51 Desai R. Propensity score fine stratification SAS macro. doi:10.7910/DVN/U8JLCW. V5 ed: Harvard Dataverse, 2020.
52 Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol* 2005;34:1370-6. doi:10.1093/ije/dyi184
53 Lash TL, Fox MP, Cooney D, Lu Y, Forshee RA. Quantitative bias analysis in regulatory settings. *Am J Public Health* 2016;106:1227-30. doi:10.2105/AJPH.2016.303199
54 Arnold BF, Ercumen A, Benjamin-Chung J, Colford JMJr. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology* 2016;27:637-41. doi:10.1097/EDE.0000000000000504
55 Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;21:383-8. doi:10.1097/EDE.0b013e3181d61eeb
56 Khosrow-Khavar F, Kim SC, Lee H, Lee SB, Desai RJ. Tofacitinib and risk of cardiovascular outcomes: results from the Safety of TofAcitinib in Routine care patients with Rheumatoid Arthritis (STAR-RA) study. *Ann Rheum Dis* 2022;81:798-804. doi:10.1136/annrheumdis-2021-221915
57 Matthews AA, Dahabreh IJ, Fröbert O, et al. Benchmarking observational analyses before using them to address questions trials do not answer: an application to coronary thrombus aspiration. *Am J Epidemiol* 2022;191:1652-65. doi:10.1093/aje/kwac098
58 Flanders WD, Khoury MJ. Indirect assessment of confounding: graphic description and limits on effect of adjusting for covariates. *Epidemiology* 1990;1:239-46. doi:10.1097/00001648-199005000-00010
59 Mor A, Dekkers OM, Nielsen JS, Beck-Nielsen H, Sørensen HT, Thomsen RW. Impact of glycemic control on risk of infections in patients with type 2 diabetes: a population-based cohort study. *Am J Epidemiol* 2017;186:227-36. doi:10.1093/aje/kwx049
60 Zelniker TA, Wiviott SD, Raz I, et al. SGLT2 inhibitors for primary and secondary prevention of cardiovascular and renal outcomes in type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. *Lancet* 2019;393:31-9. doi:10.1016/S0140-6736(18)32590-X
61 Vaduganathan M, Docherty KF, Claggett BL, et al. SGLT-2 inhibitors in patients with heart failure: a comprehensive meta-analysis of five randomised controlled trials. *Lancet* 2022;400:757-67. doi:10.1016/S0140-6736(22)01429-5
62 Qiu M, Ding L-L, Zhang M, Zhou HR. Safety of four SGLT2 inhibitors in three chronic diseases: A meta-analysis of large randomized trials of SGLT2 inhibitors. *Diab Vasc Dis Res* 2021;18:14791641211011016. doi:10.1177/14791641211011016
63 Dave CV, Schneeweiss S, Patorno E. Comparative risk of genital infections associated with sodium-glucose co-transporter-2 inhibitors. *Diabetes Obes Metab* 2019;21:434-8. doi:10.1111/dom.13531

Web appendix 1: Appendix figures and tables
Web appendix 2: Study protocol of case example study
Web appendix 3: R function for step 3a (assess expected precision)