PROTOCOL: Evaluating Reproducibility and Robustness of Real World Evidence from Large Healthcare Databases

Drafted: June 6, 2017

Abstract

Our project will produce empirically based recommendations on how to measure, document and achieve fully reproducible and robust findings from healthcare database studies. We will start by attempting to directly replicate a random sample of 250 large healthcare database studies published in a leading clinical or epidemiology journal within the last 5 years. During this effort, we will measure how frequently the reproducing team is forced to make educated guesses due to insufficient reporting of various methodology or parameter choices applied by the original investigators. We will then apply a range of plausible alternative design and algorithm choices and vary assumptions regarding confounding and misclassification to evaluate the robustness of the original study's findings.

We will identify 1) areas where insufficient detail in reporting on scientific decisions are most common, 2) operational parameters where insufficient reporting has the greatest impact on findings (e.g. "vibration ratio"[1]), and 3) predictors of insufficient reporting. Recommendations on transparency and reporting will be developed collaboratively with a Scientific Advisory Board and closely aligned with professional society guidance and journal editor policies.

Objectives for the proposed project are to:

- Measure the current state of reproducibility and robustness of healthcare database studies,
- Highlight the specific areas that most need improvement and
- Propose specific, empirically based recommendations to improve the conduct and quality of real world evidence generated from research using healthcare databases

Specific Aims

Aim 1. To quantify the current state of reproducibility via direct replication of healthcare database studies

> **Reproducibility (direct replication)** in healthcare database studies is achieved when independent investigators are able to recreate analytic cohorts and obtain the same findings by applying the same design and operational decisions to the same large healthcare data source.

Hypothesis: A substantial proportion of cohort characteristics and measures of association in published healthcare database studies will not be able to be reproduced with a reasonable margin of error.

We will systematically sample 250 healthcare database studies recently published in leading medical or epidemiologic journals, and assess how reproducible the studies' findings are by attempting to recreate the original data extraction and analysis based on publically reported methods in the papers or online appendices.

Aim 2. To evaluate the robustness of evidence currently found in healthcare database studies

> **Robustness** in healthcare database studies is the variability in study findings when study parameters are modified. It is assessed by making alternative plausible design choices (conceptual replication), changing assumptions regarding unmeasured confounding[2], quantitative and probabilistic bias[3,4] as well as use of positive and negative control outcomes.

Hypothesis 1: In studies that can be closely reproduced, findings will vary in their robustness when alternative plausible design and analysis choices are implemented.

Hypothesis 2: In studies that have clear design flaws, the findings and implications of those findings will be very different (even reversed) after those shortcomings are corrected.

We will select a targeted sample of studies from the pool of reproduced studies from Aim 1 and assess the robustness of findings, in terms of effect size and variability, to changes in assumptions regarding bias and confounding as well as plausible alternative parameter choices. We will redo the selected studies multiple times, each time changing one or more of the design and analysis parameters reported in the original papers. For studies with no obvious design or analysis flaws, we will choose plausible alternatives and assess how much this influences the resulting measures of association. For studies that have clear design or analysis flaws, we will conduct the study with the flaws corrected, use multiple plausible alternative design and analysis choices, and then compare the original and design flaw corrected measures of association.

Methods

1. Systematic search

   We will perform a systematic search using Google Scholar. The search will involve journals that are 1) highly ranked according to the h-5 index in the Health & Medical Sciences or Epidemiology subcategories in Google Scholar or 2) affiliated with the International Society of Pharmacoepidemiology or the International Society for Pharmacoeconomics and Outcomes Research.

   For each journal of interest, we will searched for results that included both of the words "cohort" and "claims," with at least one of the following words contained anywhere within the article: "Optum", "UnitedHealth", "Marketscan," "Truven" , "GPRD", "CPRD", "Medicare." The search will be limited to results published between Jan 1, 2011- Dec 31, 2016.

The sets of search results for each journal will be reviewed by members of the research team to determine whether each article qualified for inclusion in our study. These criteria included:

- *Data source mismatch*: The published study must be conducted in a database for which we have a data use agreement/license. This includes Medicare, Truven/MarketScan, United/Optum, CPRD and any combination of these databases. If the study used a database, registry, or electronic health records from a source that do not match those listed it will be excluded. If the study did not use a large healthcare database, e.g. randomized control trial or animal study, it will be excluded.
- *Year Range*: We will exclude published studies that involved years of data from the data source that are not included in our license/data use agreement.
- *Full article unavailable*: The full articles must be available for each of the reviewed papers. If the search result refers to a poster or conference abstract or members of the review team are unable to access a PDF for a full manuscript, the study will be excluded.
- *Not a descriptive or comparative safety/effectiveness cohort study*: We required that articles included in this study be a descriptive study or a comparative safety/effectiveness analysis. We will exclude articles that do not fall into this categories, including:
    - Cost effectiveness analyses
    - Methods papers (e.g. chart review validation, simulation, machine learning algorithm development)
    - Review/Meta-Analysis
    - Letter/Commentary/Editorial/Guidelines

2. Random sample

   Out of the eligible studies identified from our search, we will take a random sample of 250 studies, half descriptive and half comparative safety/efficacy.

3. Standardized extraction forms

   We will create standardized extraction forms based on the key reporting table in an ISPE/ISPOR Joint Task Force paper: Reporting to improve reproducibility and facilitate assessment of validity of healthcare database analyses v1.0 (under review). This will allow the study team to evaluate transparency of reporting on a catalogue of specific scientific decisions made by the original investigators.

4. Extract study protocols and replicate

   Study team members will review publications, appendices and cited public materials to extract technical protocols for each publication. If the publically available materials are unclear on an operational or methodologic parameter (e.g. timing of cohort entry and follow up, inclusion/exclusion criteria, algorithms to measure exposure, outcome, covariates etc.), we will record that there was insufficient detail in that area.

The technical protocol extracted from reported methods for the publication will be used to replicate original study using the same data source. When there was insufficient detail regarding a scientific decision, the team will apply a plausible alternative - the most commonly used decision or algorithm reported in other sampled papers.

5. Replication metrics

The results will include:

- Descriptive frequencies of how often reporting was insufficient for specific parameters; measuring the extent to which the study team had to make assumptions about scientific decisions made by the original investigator.
- Standardized differences between original paper and replication for prevalence or mean of baseline characteristics, incidence rates/risks, and reported measures of association (absolute and/or relative).
- Calibration plot for reported measures of association between the original and replication with bars indicating width of 95% interval for each.
- The degree to which lack of transparency in different areas (e.g. timing of cohort entry and follow up, inclusion/exclusion criteria, algorithms to measure exposure, outcome, covariates etc.) relates to standardized differences for reported measures of association in the original versus the replication.

6. Select studies for robustness checking

We will select a targeted subset of 10 studies selected where half will be studies where the original study population and findings were closely replicated and the other half will be comprised of studies where clear design flaws were identified during replication. We will assess the robustness of findings when we use plausible alternative parameter choices, make changes in assumptions regarding bias and confounding as well as evaluate negative control exposures and outcomes. Each selected study will be implemented multiple times, each time changing one or more of the items listed in Table 1.

Table 1. Robustness check comparisons

Original
Direct replication
Plausible alternative parameters
Study entry date selected before/after inclusion/exclusion criteria applied
Other algorithms used for same clinical concept for outcome
Other algorithms used for same clinical concept for covariates
Other algorithms used for same clinical concept for inclusion/exclusion
Different stockpiling/exposure risk criteria
Different washout criteria
Different censoring criteria
Different years of data in same data source
Different data source
Correction of clear design flaws
Combinations of the variations above
External adjustment
Residual confounding
Quantitative and probabilistic bias

Negative control outcome/exposure

7. Robustness metrics

   The results will include:

   - Proportion of full sample of 250 studies with clear design or analysis flaws stratified by type. Design flaws may include immortal time bias, reverse causation, adjustment for intermediates, use of inappropriate comparators, etc. Detailed rationale behind assessment of flawed design choices will be provided in supporting materials.
   - Vibration ratio[1]: The degree to which effect sizes varied with alternative parameter choices (largest measure of association/smallest measure of association)
   - Figure plotting measures of association with 95% confidence intervals for 1) original paper, 2) replication using the originally reported methods, 3) replications using a set of plausible alternative choices, 4) after making assumptions regarding residual confounding, 5) after quantitative and probabilistic bias correction, 6) correction of clear design flaws
   - Figure plotting measures of association with 95% confidence intervals for 1) original paper, 2) replication using the originally reported methods, 3) negative control exposure and outcomes

8. Collaboration with original authors

   After replicating the 250 randomly sampled studies, we will reach out to the corresponding authors on each paper to discuss the specific operational decisions and assumptions made when we attempted to directly replicate their study. For original authors who are interested in collaborating, we will work together to understand differences between the original study and our replication conducted using the same years of data from the same large healthcare data source. The collaboration will focus on robustness of findings and moving toward greater clarity in reporting of decisions made during study implementation.

9. Scientific Advisory Board

   A Scientific Advisory Board comprised of international representatives from regulatory and health technology assessment agencies, journal editors, payer organizations, large healthcare database networks, professional research societies, industry and patients will be involved with this project from the beginning. A critical component of the ability of this project to provide relevant and actionable evidence to shape policy and guidelines is the early and continued engagement of key stakeholders.

10. Transparency in evaluation of reproducibility and robustness of real world evidence from large healthcare databases

We will create and maintain a website for this project and related efforts.

Items available on the website before replications are complete:

- Study protocol
- Link to ENCePP registration
- ISPOR/ISPE joint task force papers
- Other efforts…?

Items available on the website after replications complete:

- A full list of the 250 randomly sampled studies
- A complete report of study parameters used for each replication
- Tables comparing baseline characteristics and measures of association for the original paper and the replication
- A key/complete list of parameters varied in robustness checks for each selected study
- Proportion of original authors who responded to study team's questions about published paper and distribution of types of responses
- PDF of papers published from replication/robustness projects

References:

1. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology.* Sep 2008;19(5):640-648.

2. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and drug safety.* May 2006;15(5):291-303.

3. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *International journal of epidemiology.* Jul 30 2014.

4. Lash TL, Vandenbroucke JP. Should preregistration of epidemiologic study protocols become compulsory? Reflections and a counterproposal. *Epidemiology.* Mar 2012;23(2):184-188.