Prespecified Falsification End Points
Can They Validate True Observational Associations?

Vinay Prasad, MD
Anupam B. Jena, MD, PhD

As observational studies have increased in number—fueled by a boom in electronic recordkeeping and the ease with which observational analyses of large databases can be performed—so too have failures to confirm initial research findings.[1] Several solutions to the problem of incorrect observational results have been suggested,[1,2] emphasizing the importance of a record not only of significant findings but of all analyses conducted.[2]

An important and increasingly familiar type of observational study is the identification of rare adverse effects (defined by organizations such as the Council for International Organizations and Medical Sciences as occurring among fewer than 1 per 1000 individuals) from population data. Examples of these studies include whether macrolide antibiotics such as azithromycin are associated with higher rates of sudden cardiac death[3]; whether proton pump inhibitors (PPIs) are associated with higher rates of pneumonia[4]; or whether bisphosphonates are associated with an increased risk of atypical (subtrochanteric) femur fractures.[5] Rare adverse events, such as these examples, occur so infrequently that almost by definition they may not be identified in randomized controlled trials (RCTs). Postmarketing data from thousands of patients are required to identify such low-frequency events. In fact, the ability to conduct postmarketing surveillance of large databases has been heralded as a vital step in ensuring the safe dissemination of medical treatments after clinical trials (phase 4) for precisely this reason.

Few dispute the importance of observational studies for capturing rare adverse events. For instance, in early studies of whether bisphosphonate use increases the rate of atypical femur fractures, pooled analysis of RCTs demonstrated no elevated risk.[6] However, these data were based on a limited sample of 14 000 patients with only 284 hip or femur fractures and only 12 atypical fracture events over just more than 3.5 years of follow-up. In contrast, later observational studies addressing the same question were able to leverage much larger and more comprehensive data. One analysis that examined 205 466 women who took bisphosphonates for an average of 4 years identified more than 10 000 hip or femur fractures and 716 atypical fractures.[5] This analysis demonstrated an increased risk of atypical fractures associated with bisphosphonate use and was validated by another large population-based study.

However, analyses in large data sets are not necessarily correct simply because they are larger. Control groups might not eliminate potential confounders, or many varying definitions of exposure to the agent may be tested (alternative thresholds for dose or duration of a drug)—a form of multiple-hypothesis testing.[2] Just as small, true signals can be identified by these analyses, so too can small, erroneous associations. For instance, several observational studies have found an association between use of PPIs and development of pneumonia, and it is biologically plausible that elevated gastric pH may engender bacterial colonization.[4] However, it is also possible that even after statistical adjustment for known comorbid conditions, PPI users may have other unobserved health characteristics (such as poor health literacy or adherence) that could increase their rates of pneumonia, apart from use of the drug. Alternatively, physicians who are more likely to prescribe PPIs to their patients also may be more likely to diagnose their patients with pneumonia in the appropriate clinical setting. Both mechanisms would suggest that the observational association between PPI use and pneumonia is confounded. In light of the increasing prevalence of such studies and their importance in shaping clinical decisions, it is important to know that the associations identified are true rather than spurious correlations. Prespecified falsification hypotheses may provide an intuitive and useful safeguard when observational data are used to find rare harms.

A falsification hypothesis is a claim, distinct from the one being tested, that researchers believe is highly unlikely to be causally related to the intervention in question.[7] For instance, a falsification hypothesis may be that PPI use increases the rate of soft tissue infection or myocardial infarction. A confirmed falsification test—in this case, a positive association between PPI use and risks of these conditions—

Author Affiliations: Medical Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland (Dr Prasad); Department of Health Care Policy, Harvard Medical School, and Massachusetts General Hospital, Boston (Dr Jena); and National Bureau of Economic Research, Cambridge, Massachusetts (Dr Jena).
Corresponding Author: Anupam B. Jena, MD, PhD, Department of Health Care Policy, Harvard Medical School, 180 Longwood Ave, Boston, MA 02115 (jena@hcp.med.harvard.edu).

would suggest that an association between PPI use and pneumonia initially suspected to be causal is perhaps confounded by unobserved patient or physician characteristics. Ideally, several prespecified false hypotheses can be tested and, if found not to exist, can support the main study association of interest. In the case of PPIs, falsification analyses have shown that many improbable conditions—chest pain, urinary tract infections, osteoarthritis, rheumatoid arthritis flares, and deep venous thrombosis—are also linked to PPI use,[4] making the claim of an increased risk of pneumonia related to use of the drug unlikely.

Another example of falsification analysis applied to observational associations involves the reported relationship of social networks with the spread of complex phenomena such as smoking, obesity, and depression. In social network studies, persons with social ties are shown to be more likely to gain or lose weight, or to start or stop smoking, at similar time points than 2 random persons in the same group. Several studies supported these claims; however, other studies have shown that even implausible factors—acne, height, and headaches—may also exhibit "network effects."[8]

Falsification analysis can be operationalized by asking investigators to specify implausible hypotheses up front and then testing those claims using statistical methods similar to those used in the primary analysis. Falsification could be required both for studies that aim to show a rare harm of a particular medical intervention as well as for studies that aim to show deleterious interactions between medications. For instance, in evaluating whether concomitant use of clopidogrel and PPIs is associated with decreased effectiveness of the former drug and worsens cardiovascular outcomes, does the use of PPIs also implausibly diminish the effect of antihypertensive agents or metformin?

Prespecifying falsification end points and choosing them appropriately is important for avoiding the problem of multiple hypothesis testing. For instance, if many falsification hypotheses are tested to support a particular observational association, a few falsification outcomes will pass the falsification test—ie, will not be associated with the drug or intervention of interest—whereas other falsification tests may fail. If the former are selectively reported, some associations may be mistakenly validated. This issue cannot be addressed by statistical testing for multiple hypotheses alone because selective reporting may still occur. Instead, prespecifying falsification outcomes and choosing outcomes that are common may mitigate concerns about post hoc data mining. In the case of PPIs and risk of pneumonia, falsification analyses used prevalent ambulatory complaints such chest pain, urinary tract infections, and osteoarthritis.[4]

Observational studies of rare effects of a drug may be further validated by verification analyses that demonstrate the presence of known adverse effects of a drug in the data set being studied. For instance, an observational study suggesting an unknown adverse effect of clopidogrel (for example, seizures) should also be able to demonstrate the presence of known adverse effects such as gastrointestinal hemorrhage associated with clopidogrel use. The inability of a study to verify known adverse effects should raise questions about selection in the study population.

Although no published recommendations exist, standardized falsification analyses with 3 to 4 prespecified or highly prevalent disease outcomes may help to strengthen the validity of observational studies, as could inclusion of verification analyses. Information on whether falsification and validation end points were used in a study should be included in a registry for observational studies that others have suggested.[2]

Prespecified falsification hypotheses can improve the validity of studies finding rare harms when researchers cannot determine answers to these questions from RCTs, either because of limited sample sizes or limited follow-up. However, falsification analysis is not a perfect tool for validating the associations in observational studies, nor is it intended to be. The absence of implausible falsification hypotheses does not imply that the primary association of interest is causal, nor does their presence guarantee that real relations do not exist. However, when many false relationships are present, caution is warranted in the interpretation of study findings.

Conflict of Interest Disclosures: The authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest and none were reported.

REFERENCES

1. Thomas L, Peterson ED. The value of statistical analysis plans in observational research: defining high-quality research from the start. *JAMA*. 2012;308(8):773-774.
2. Ioannidis JP. The importance of potential studies that have not existed and registration of observational data sets. *JAMA*. 2012;308(6):575-576.
3. Ray WA, Murray KT, Hall K, Arbogast PG, Stein CM. Azithromycin and the risk of cardiovascular death. *N Engl J Med*. 2012;366(20):1881-1890.
4. Jena AB, Sun E, Goldman DP. Confounding in the association of proton pump inhibitor use with risk of community-acquired pneumonia [published online September 7, 2012]. *J Gen Intern Med*. doi:10.1007/s11606-012-2211-5.
5. Park-Wyllie LY, Mamdani MM, Juurlink DN, et al. Bisphosphonate use and the risk of subtrochanteric or femoral shaft fractures in older women. *JAMA*. 2011;305(8):783-789.
6. Black DM, Kelly MP, Genant HK, et al; Fracture Intervention Trial Steering Committee; HORIZON Pivotal Fracture Trial Steering Committee. Bisphosphonates and fractures of the subtrochanteric or diaphyseal femur. *N Engl J Med*. 2010;362(19):1761-1771.
7. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Q J Econ*. 2004;119:249-275.
8. Cohen-Cole E, Fletcher JM. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ*. 2008;337:a2533.