Young PointOfView

GeneticEngineering & Biotechnology News May 1, 2014 (Vol. 34, No. 9)

http://www.genengnews.com/gen-articles/are-medical-articles-true-on-health-disease/5203/?kwrd=young

Are Medical Articles True on Health, Disease?

Sadly, Not as Often as You Might Think

S. Stanley Young

Henry I. Miller, M.D.

Science works only when experiments are reproducible. If an experiment cannot be replicated, both the scientific enterprise and those who depend upon its results are in trouble. Driven by the realization that experiments surprisingly often do not replicate, the issue of claims in scientific papers is receiving increasing scrutiny. Given that biomedical research is one of the most important goals of the scientific enterprise, it is especially important to know how well the claims that result from clinical studies hold up.

Table

1	Design of experiments
2	Data construction, moving from raw data to an analysis file
3	Simple data-handling mistakes
4	Multiple testing
5	Multiple modeling
6	Bias in observational studies due to imbalanced covariates
7	A p-value of <0.05 is not strong enough
8	Publication bias
9	Fraud
10	Inadequate scientific oversight
11	Perverse incentives

The published literature on observational studies—which are noninterventional but rely on data mining, the searching of large datasets for patterns of association—is notoriously unreliable. In randomized clinical trials (RCTs), which are commonly viewed as the gold standard for accuracy and reliability, observational studies are poorly replicated.

Young and Karr[1] found 12 articles in prominent journals in which 52 claims coming from observational studies were tested in randomized clinical trials. Many of the RCTs were quite large, and most were run in factorial designs, e.g., vitamin D and calcium individually and together, along with a placebo group. Remarkably, none of the claims replicated in the direction claimed in the observational studies; in five instances there was actually statistical significance in the opposite direction.

After a series of failed attempts to extend basic research findings (from academic labs), two large drug companies, Bayer and Amgen, carefully reviewed their experience and found that only 25 and 11 percent, respectively, of the claims in the scientific literature could be replicated in a way that was sufficiently robust to be useful as the basis for drug development projects.[2,3] Astonishingly, even when the investigators asked the original researchers to replicate their own work, for the most part they could not. This may help to explain the difficulty of translating cancer research in the laboratory to clinical success.

Ioannidis[4] provides insight into the reproducibility of medical claims in the literature. He analyzed how well claims in highly cited papers in major medical journals could be replicated and found a large discrepancy in success rates between RCTs and observational studies—19 of 28 (67.9%) versus 1 of 6 (16.6%), respectively.

Replication rates of 0.0%,[1] 67.9%, or 16.6%,[4] are unacceptable. Clearly the standard p-value of <0.05 as a measure of statistical significance is not a reliable indicator that a result will replicate. It is worth enumerating some of the problems that can lead to a claim failing to replicate (*Table*).

Claims that Fail to Replicate

Some of these factors are obvious, while others need explication. For example, Item 1: RCTs are usually carefully designed and executed, particularly if corporate money is at risk and the FDA is involved. But not always: The failures of design include insufficient statistical power (too few patients to reliably detect a likely effect); inappropriate choice of route, dose or frequency of administration, or in the stratification of subjects.

Observational studies need to be much better designed to more closely resemble RCTs, even though they cannot be prospectively randomized. Indeed, some observational datasets may not be suitable for any analysis.

Item 2: In observational studies, usually some raw data must be manipulated to get them into a form suitable for analysis. This can be complex and, when combined with design choice (Item 1), the way it is done can dramatically change analysis results.[5]

Items 4, 5, 10 and 11: Journals generally require a p-value <0.05 to merit consideration for publication, but because they do not require investigators to make datasets available, there may be an incentive to manipulate the process to get a p-value that "qualifies." There is general agreement that fabricating data is fraud, but is it legitimate to ask hundreds of questions and/or look at thousands of models and not show how these choices affect the resulting p-values and claims?

Item 5: "Multiple modeling" provides the analyst with the flexibility to manipulate the data in order to get a p-value <0.05. For example, the analysis can be adjusted using linear models to make treatment groups more similar; with 10 covariates there are 1024 possible ways to perform this adjustment. Patient matching can be done in a number of ways. But such subtleties are largely hidden from the reader.

Item 7: There have been claims that using a p-value cutoff at 0.05 is not sufficiently stringent.[6] RCTs used to support drug approval require two studies with p-value <0.05 for an effective p-value ~0.0025. Industry-funded RCTs replicate with a frequency of about 78.6%, while other RCTs, typically using a single p-value <0.05, replicate 57.1% of the time. It should be noted that if the studies used to replicate the original study were powered at the commonly used 80%, then it is possible that all of the industry claims and some of the other RCT studies are correct, i.e., the replicate study could be wrong.

Items 9 and 11: The world record-holder for fraudulent papers may be Yoshitaka Fujii, who published 212 papers over about 20 years, only three of which were clearly not fraudulent. Had he been required to deposit his datasets, his misconduct might have been discovered sooner.

Observational Studies

There appear to be systemic problems with the way that observational studies are commonly conducted. Virtually all of the problems listed in the *Table* can plague observational studies and, of course, any one alone or a combination of them could wreck a study. In light of multiple testing and multiple modeling, a p-value <0.05 is not nearly rigorous enough.[6] Five of the six observational studies in Ioannidis[4] reported p-values of 0.0001, 0.001, 0.003, 0.008, and 0.015 (the 6th study was a case series and reported no p-value). Most of these p-values are small enough that they would be considered either "strong" or even "decisive" evidence by statistics professor Valen E. Johnson,[6] but in all these cases, well-designed RCTs failed to confirm the claims made in these observational studies.

It is popular to blame investigators for these problems, but the culpability must be shared by the managers of the scientific process: funding agencies and journal editors. At a minimum, funding agencies should require that datasets used in papers be deposited so that the normal scientific peer oversight can occur. Journal editors need to reexamine their policy of being satisfied with a p-value <0.05, unadjusted for multiple testing or multiple modeling. Editors are using "quality by inspection" (p-value <0.05) rather than the more modern "quality by design."

Joiner and Gaudard (Qual. Progr., 1990) contrasted the difference between quality by inspection and quality by design: "Depending on inspection is like treating a symptom while the disease is killing you. The need for inspection results from excessive variability in the process. . . Ceasing dependence on variability means you must understand your processes so well that you can predict the quality of their output from upstream measurements and activities."

If a question is important and the study is well-designed and well-conducted, then the work should be publishable. By essentially requiring a p-value <0.05, editors are directly responsible for publication bias, because most negative studies are not published.

Finally, where does all this leave the consumers of medical studies? For the interpretation of randomized clinical trials that use a single p-value <0.05, it is *caveat lector*. For observational studies, it would seem that replication should be a bare minimum before one credits the result; and even replication may not be enough.

Obviously, the treatment of data in medical studies has a long way to go.

S. Stanley Young (young@niss.org) is the assistant director for bioinformatics at the National Institute of Statistical Sciences. Henry I. Miller (henry.miller@stanford.edu) is the Robert Wesson Fellow in Scientific Philosophy and Public Policy at Stanford University's Hoover Institution.

References:

[1] Young SS, Karr A. Deming, data and observational studies: A process out of control and needing fixing. Significance 2011;September:122–126.
[2] Prinz F, SchlangeT, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Rev. Drug Discov. 2011;10:712-713.
[3] Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature 2012;483:531-533.
[4] Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA 2005;294:218–229.
[5] Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. Therapeutic Advances in Drug Safety. 2013;4:53-62.
[6] Johnson VE. Revised standards for statistical evidence. PNAS 2013;110:19313-19317.