Research integrity, and why bad science in biomedicine and agriculture has become such a problem

Science depends on corroboration — that is, researchers verify others' results, often making incremental advances as they do so. The nature of science dictates that no research paper is ever considered to be the final word, but increasingly, there are too many whose results are not reproducible. Explanations include the complexity of experimental systems, misunderstanding (and often, misuse) of statistics, pressures on researchers to publish, and the proliferation of shoddy pay-to-play "predatory" journals.

In 2011 and 2012, two articles rocked the scientific world. One reported the attempt to reproduce the results of 53 preclinical research papers that were considered "landmark" studies. The scientific findings were confirmed in only six (11%) of them. Astonishingly, even the researchers making the claims *could not replicate their own work*.

The second article found that claims made using observational data could not be replicated in randomized clinical trials (which is why the latter are known as the "gold standard). Overall, there were 52 claims tested and *none* replicated in the expected direction, although most had very strong statistical support in the original papers.

Subsequently, there has been more evidence of a crisis in scientific research: In a survey of ~1500 scientists, 90% said there were major or minor problems with the replication of experiments.

More recently, in 2015, 270 co-investigators published the results of their systematic attempt to replicate work reported in 98 original papers from three psychology journals, to see how their results would compare. According to the replicators' qualitative assessments, only 39 of the 100 replication attempts were successful.

Around the same time, a multinational group attempted to replicate 21 systematically selected experimental studies in the social sciences published in the journals *Nature* and *Science* between 2010 and 2015. They found "a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size."

These failure rates for reports in prominent journals are astonishing — and worrisome, because false claims can become canonized.

Of course, technical problems with laboratory experiments – contamination of cell lines or reagents; unreliable equipment; the difficulty of doing a complex, multi-step experiment the same way, time after time; etc. – are one explanation, but another is statistical sleight-of-hand. One technique for that is called p-hacking: Scientists try one statistical or data manipulation after another until they get a small p-value that qualifies as "statistical significance," although the finding is the result of chance, not reality.

Australian researchers examined all the publicly available literature and found evidence that p-hacking was common in almost every scientific field. Peer review and editorial oversight are inadequate to ensure

that articles in scientific publications represent reality instead of statistical chicanery. Another problem is that competing scientists often do not retest questions, or if they do, they don't make known their failure to replicate, so there are significant lacunae, or gaps, in the published literature – which, of course, mostly comes from universities and is funded by taxpayers.

Many claims appearing in the literature do, of course, replicate, but even those may not be reliable. Many claims in the psychology literature, for example, are only "indirectly" replicated. If X is true, then Y, a consequence, should also be true. Often Y is accepted as correct, but it turns out that neither X nor Y replicates when tested anew.

Understandably, editors and referees are biased against papers that report negative results; they greatly prefer positive, statistically significant results. Researchers know this and often don't even submit them – the so-called "file drawer effect." Once enough nominally positive, confirmatory papers appear, the claim becomes canonized, making it even more difficult to publish an article that reports a contrary result.

The system thus perverts the method, the value of accumulated data, and the dogma of science. It makes us wonder whether scientists who practice statistical trickery fail to understand statistics, or whether they're so confident of the correct outcome that they take shortcuts to get to it. If the latter, it would bring to mind the memorable observation about science by the late, great physicist and science communicator Richard Feynman, "The first principle is that you must not fool yourself – and you are the easiest person to fool."

Part of the canonization process often involves a meta-analysis, which is defined as "a method for systematically combining pertinent qualitative and quantitative study data from several selected studies to develop a single conclusion that has greater statistical power [and that] is statistically stronger than the analysis of any single study, due to increased numbers of subjects, greater diversity among subjects, or accumulated effects and results."

This is how it's done… A computer search finds published articles that address a particular question – say, whether taking large amounts of vitamin C prevents colds. From those that are considered to be methodologically sound, the data are consolidated and carried over to the meta-analysis. If the weight of evidence, based on a very stylized analysis, favors the claim, it is determined to be real, or canonized.

The problem is that there may *not* be safety in numbers because many of the individual base papers are very likely wrong – the result of p-hacking and publication bias. Potential p-hacking can be detected by creating a "p-curve" – i.e., plotting the p-values for each of the papers included in the meta-analysis against the "rank" — the integers 1,2,3…etc., up to the number of papers. The first figure below, for example, plots a meta-analysis in which there were 19 papers; in the second figure, the meta-analysis included 14 papers.

Recall that a p-value measures the likelihood that an effect is real, as opposed to having occurred by chance. The smaller the p-value, the more likely the effect is real.

If the resulting p-curve looks like a hockey stick, with small p-values on the blade and larger p-values on the handle (as in the two figures below), there is a good case to be made for p-hacking.

The figures below are derived from meta-analyses of the supposedly beneficial effects of omega-3 fatty acids and the alleged direct relationship between sulfur dioxide in the air and mortality, respectively that were presented in a major medical journal and claimed a positive effect. There are, indeed, several small p-values reported and, taken alone, they would indicate a real effect. But there are more p-values greater than 0.05, which indicate no effect. Both cannot be correct. Inasmuch as there are many more negative studies and p-hacking is the logical explanation for the presence of a small number of low p-values, the most likely conclusion is that there is no effect. Thus, the meta-analyses yield false-positive results.

These examples are all too common. The sad truth is that much of published science and the canonized claims resulting from it are likely wrong, and it is incumbent on the scientific community to find solutions. Without research integrity, we don't know what we know.

unnamed file

Image not found or type unknown

Dr. S. Stanley Young is a statistician who has worked at pharmaceutical companies and the National Institute of Statistical Sciences on questions of applied statistics. He is an adjunct professor at several universities and a member of the EPA's Science Advisory Board

Henry I. Miller, a physician and molecular biologist, is a Senior Fellow at the Pacific Research Institute in San Francisco. He was the founding director of the FDA's Office of Biotechnology. Follow him on Twitter @henryimiller