WAYNE STATE School of Medicine | **Department of Biochemistry, Microbiology and Immunology**

BMI Teaching Trends Newsletter – October 2021

Teaching Matters
By Dr. Matthew Jackson

Misleading Scientific Research

The COVID-19 pandemic has renewed interest in the integrity of research with particular emphasis on clinical trials that are aimed at demonstrating the effectiveness of therapeutic interventions. Widely popularized pseudo-therapies such as the antiparasitic drugs hydroxychloroquine and ivermectin have revealed the glaring disconnect between public insight and reality. Unfortunately, the past few years have forced us to accept that practicable reasoning skills are beyond the capability of a large segment of the population. However, the question is: how discriminating is the scientific community when it comes to recognizing bad science? As described in a June 2020 article from Inside Higher Ed, the rush toward COVID-19 research has exposed weaknesses in the scholarly publication system. Research that has appeared in preprint servers such as arXiv, bioRxiv, and SSRN attract attention on social media as well as from quasi-legitimate news agencies. Upon dissemination, the lay public and non-evidence-based practitioners may regard weak science as fact ultimately leading to unsubstantiated treatment decisions with potentially fatal consequences. In an effort to provide peer-reviewed research findings without the months-long review process, MIT Press and the Berkeley School of Public Health are launching a new journal with the goal of minimizing the negative impact that some of the preprint servers have created. *Rapid Reviews: COVID-19* is aimed at elevating legitimate research reporting in a timely manner thereby facilitating research advances based on peer-reviewed findings.

COVID-19 research has focused attention on scientific research as a whole. A recent article in the libertarian magazine, Reason titled *How Much Scientific Research Is Actually Fraudulent?* summarized ongoing concerns, particularly with sloppy than randomized drug trials, and reaffirmed the expectation that all datasets should be made available for reanalysis by other researchers. That article describes the too common practice of HARKing (Hypothesizing After the Results are Known) which is defined as presenting a post hoc hypothesis based on the investigator's results as an a priori hypotheses. *Saving Science* is an earlier review by Daniel Sarewitz that provides some staggering statistics from various fields of biomedical research: i) in 2012, a major biotechnology company found that only six out of 53 landmark published preclinical cancer studies could be replicated; ii) when amyotrophic lateral sclerosis researchers tested more than 100 potential drugs reported to slow disease progression in mouse models, none were found to be beneficial when tested on the same mouse strains; iii) 1,000 peer-reviewed and published

breast cancer research studies turned out to be using a skin cancer cell line; iv) fMRI brain imaging studies suffered from a 70 percent false positive rate; and v) decades of nutritional dogma about the alleged health dangers of salt, fats, and red meat appears to be wrong. A root cause of the poor science stemmed from the lack of reproducibility which is a hallmark of hypothesis-driven research.

A 2005 report by John P. A. Ioannidis, *Why Most Published Research Findings Are False* focuses on the high rate of nonreplication and poor statistical analyses as a basis for bad science. This author points out that too often research findings are simply accurate measures of the prevailing bias are therefore do little to advance the field. A strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05, leads to the lack of reproducibility. Research findings are less likely to be true: i) when the sample size is small; ii) when there is a small power of significance; iii) when there is greater flexibility in study design and outcome measures (i.e., bias); iv) when there are greater financial incentives; and, v) when the field is hot and the team involved is extraordinarily large. Ioannidis emphasizes the importance of pre-study odds in the following example:

Box 1. An Example: Science at Low Pre-Study Odds

Let us assume that a team of investigators performs a whole genome association study to test whether any of 100,000 gene polymorphisms are associated with susceptibility to schizophrenia. Based on what we know about the extent of heritability of the disease, it is reasonable to expect that probably around ten gene polymorphisms among those tested would be truly associated with schizophrenia, with relatively similar odds ratios around 1.3 for the ten or so polymorphisms and with a fairly similar power to identify any of them. Then $R = 10/100,000 = 10^{-4}$, and the pre-study probability for any polymorphism to be associated with schizophrenia is also $R/(R + 1) = 10^{-4}$. Let us also suppose that the study has 60% power to find an association with an odds ratio of 1.3 at $\alpha = 0.05$. Then it can be estimated that if a statistically significant association is found with the p-value barely crossing the 0.05 threshold, the post-study probability that this is true increases about 12-fold compared with the pre-study probability, but it is still only 12×10^{-4}.

Now let us suppose that the investigators manipulate their design, analyses, and reporting so as to make more relationships cross the $p = 0.05$ threshold even though this would not have been crossed with a perfectly adhered to design and analysis and with perfect comprehensive reporting of the results, strictly according to the original study plan. Such manipulation could be done, for example, with serendipitous inclusion or exclusion of certain patients or controls, post hoc subgroup analyses, investigation of genetic contrasts that were not originally specified, changes in the disease or control definitions, and various combinations of selective or distorted reporting of the results. Commercially available "data mining" packages actually are proud of their ability to yield statistically significant results through data dredging. In the presence of bias with $u = 0.10$, the post-study probability that a research finding is true is only 4.4×10^{-4}. Furthermore, even in the absence of any bias, when ten independent research teams perform similar experiments around the world, if one of them finds a formally statistically significant association, the probability that the research finding is true is only 1.5×10^{-4}, hardly any higher than the probability we had before any of this extensive research was undertaken!

Advances in research into COVID-19 that is currently plaguing the human population has led to the rapid development of legitimate therapies, vaccines, and predictive modeling of virus mutations all derived from evidence-based, statistically rigorous research. It has also renewed the need for critical appraisal and the systematic weeding out of research findings that are not hypothesis-driven with reproducible results. The reviews cited here point out that the problem is more acute with clinical trials which are typically motivated by substantial financial gain with sizeable authorships. How many clinical decisions are based on HARKing and the use of statistical packages that yield the desired results through data dredging?