# Preface

There is a problem in biomedical research. Whether it is a crisis and whether it is a new problem or an old endemic problem newly recognized are open questions. Whether it represents an ominous turn of events or merely a hiccup in the self-correcting process of biomedical research also is an open question. At the very minimum, it may be just a crisis of confidence. But it seems to have captured the imagination and concern of journal editors and administrators of research-granting institutions and, consequently, should be of concern to everyone involved in biomedical research.

The problem is the failure of reproducibility of biomedical research. Indeed, as will be discussed in this book, there is an appreciation of local reproducibility in that virtually every experiment involves more than one observation or trial. Just as virtually every experiment, at least implicitly, appreciates the importance of replication within an experiment, the same issues and concerns apply to the larger issues of reproducibility across different experiments by different researchers. Even when the failure of reproducibility is defined narrowly, such as a failure to achieve the same results when other researchers independently replicate the experiment—the narrow sense of irreproducibility—there are concerns. (Note that narrow reproducibility is different from local reproducibility, described previously, and broad reproducibility, described later.) This narrow sense often focuses on issues of fraud, transparency, reagents, materials, methods, and statistical analyses for which better "policing" would solve the problem. Perhaps these may be the major factors, but they may not be the only factors. None of this is to deny the importance of fraud, transparency, reagents, materials, methods, and statistical analyses, but at the same time it is possible that some causes of irreproducibility may be in the logic inherent in the research studies. Indeed, numerous examples will be presented and, thus, there is an obligation to carefully consider the logical basis of any experiment. Also, absent from the discussion is the fact that irreproducibility in a productive sense is fundamental to scientific progress. Such consideration is the central theme of this book.

Generally, results of studies are dichotomized into "positive" and "negative" studies. Irreproducibility affects positive and negative studies differently. Most of the debate centers on positive claims later demonstrated to be false, examples of a type I error—claiming as true what

is false. In statistics, a type I error is occasioned by rejecting the null hypothesis inappropriately, which posits no difference in some measured phenomenon between the experimental and the control samples. Understandably, type I errors shake confidence and risk misdirecting subsequent research. Perhaps even more of a problem are type II errors, where negative studies claiming, falsely, that there is no change in a phenomenon as a result of an experimental manipulation or no differences between phenomena just because the null hypothesis could not be rejected. It is just possible that the experiment, by design or statistical circumstance, would have only a low probability of being able to reject the null hypothesis. A better term for these studies are "null studies," as using the term "negative" implies some inference with a degree of confidence even if undefined. The null study contrasts to situations of a negative study where the null hypothesis is rejected in a setting of sufficient statistical power and thus a positive claim of no difference can be made with confidence (equivalence, noninferiority, or nonsuperiority studies). The latter is termed a negative study and provides confidence just as the modus tollens form of the Scientific Method provides certainty, as will be discussed in the text. The extent of the problem of null studies is magnified by the bias toward not publishing null studies and, when published, are typically confused as negative studies. Type II errors in null studies may result in lost opportunities by discouraging further investigations.

Replicability, the narrow sense of reproducibility, is of primary importance as it is the first requirement. The author agrees with those who are concerned about irreproducibility in the narrow sense and indeed supports proposals for the policing of fraud; transparency; reporting of reagents, materials, and methods; and robust statistical analyses—but it does not and should not stop there. There is the temptation to dismiss concerns about irreproducibility in the narrow sense, believing it acceptably low and little more than the cost of doing scientific business. Perhaps so, if these irreproducible studies were merely outliers or flukes and that the vetting process for awarding grants and accepting papers for publication is otherwise effective. However, one must remind oneself that many of the papers describing research subsequently found to be irreproducible were vetted by experts in the field. What does this say about such expertise or the process?

There may be another view, which is the broad sense of irreproducibility, that does not focus on the exact replication of a specific experiment, but rather the failure to generalize or translate. These may be called conceptually irreproducible. Of particular concern is the failure to generalize or translate from nonhuman studies to the human condition; after all, is that not the *raison d'être* of the National Institutes of Health? But even further, is it not a founding pillar of modern science's

Reductionism where there is faith in the ability to generalize and translate from the reduced and simplified? When considered in the broader sense, there is far more evidence for and concern about these forms of irreproducibility. One need look no further than the postmarket withdrawal of drugs, biologics, and devices by organizations such as the US Federal Food and Drug Administration (FDA), whose preapproval positive clinical trials were supposedly vetted by experts and found valid. Only later were the efficacy and safety conclusions found irreproducible. This is not without consequences for patients.

The lives of biomedical researchers would be made much easier if irreproducibility was merely the result of fraud; improper use of statistics; lack of transparency; or failure to report reagents, materials, and methods. But what if there are other causes of irreproducibility fundamental to the paradigms of biomedical research, such as the Scientific Method (hypothesis-driven research) and statistics? As will be argued in this text, much of scientific progress requires the use of logical fallacies in order to gain new knowledge. Strict deduction while providing the greatest certainty in the conclusions in an important sense does not create new knowledge and induction to new knowledge is problematic. Traditional valid logical deduction is the logic of certainty, it is not the logic of discovery. Indeed, the Scientific Method is an example of the logical Fallacy of Confirming the Consequence (or Consequent), also known as the Fallacy of Affirming the Consequence (or Consequent). Also, the logic of the Scientific Method is referred to as abduction. This claim itself is not controversial as both scientists and philosophers have pointed this out for decades. What is novel here will be the demonstration that this fallacy is a cause of irreproducibility in biomedical research. Fortunately, if recognized, there are methods to blunt the effect of the fallacy, thereby reducing the risk of unproductive irreproducibility. What is needed, it will be argued, is the judicious use of logical fallacies.

Perhaps novel and counterintuitive, at least from the perspective of some scientists, is that the discipline of logic, typically in the domain of philosophy, could be considered relevant to empirical biomedical research. Scientists from the beginning of modern science, as seen in the founders of the Royal Society, rejected philosophy by taking aim against scholastic metaphysician natural philosophers. While scientists as late as the early 1900s invoked past philosophers in scientific discussions as Sir Charles Sherrington did with Rene Descartes in his text "The Integrative Action of the Nervous System" in 1906, such discussions are virtually absent in any literature typically read by today's biomedical researcher. Thus, it is understandable that biomedical researchers would be skeptical of any argument that proceeds from anything that smacks of philosophy, such as logic. But lack of experience in logic,

epistemology (concerning how knowledge is gained rather than the content of knowledge), more generally, is a poor basis for skepticism. All this author can do is ask for forbearance, as the author is confident that such patience will be rewarded.

This author's ambitions for any role of logic in this particular text are circumscribed. It is critical to appreciate that logic alone does not create biomedical scientific knowledge. Science is fundamentally empiric and its success or failure ultimately relies on observation, data, and demonstration. All logic can do is provide some degree of certainty to the experimental design and analytical methods that drive claims of new scientific knowledge. Yet, the issues of reproducibility fundamentally involve issues of certainty, as reproducibility is at its core a testament to certainty. On this basis alone, logic has a role to play in concerns about scientific reproducibility. The discussions in this text are intended to strengthen, perhaps by just a bit, the already strong and important position of empirical biomedical research.

It may well be that an experiment is so obvious that there are no concerns as to the underlying logic raised. However, the success of such an experiment is not evidence that logic is not operating. For example, consider the Human Genome Project, which has been called a descriptive research program in contrast to a hypothesis-driven program. Perhaps it could be argued that the domain of the research was clearly defined and marked, that is, the human genome, thereby obviating any inductive ambiguity. Essentially, the Human Genome Project consisted of "turning the crank." Interestingly, the project set the stage for subsequent hypothesis-driven research (Verma, 2002), for example, that identified *gene A* causes *disease B* such that affecting *gene A* cures *disease B*. In that regard, hypothesis-driven research has not fared well given only two FDA-approved gene therapies (as opposed to genetic testing), despite the estimated $3 billion spent on the Human Genome Project. It is important to note that this is not a criticism of the Human Genome Project and that there is every reason to believe that the results will change medical therapies dramatically, but it will take time because it is difficult to go from data collection to cause and effect required of a hypothetico-deductive approach critical to biomedical research.

Further attesting to the potential contributions of logic is the fundamental fallacy inherent in statistics as used in biomedical research, which is the Fallacy of Four Terms. The experimental designs typically involve hypothesis testing on a sample thought representative of the population of concern with inferences from the findings on the sample transferred to the population through a syllogistic deduction. The sample is the entities studied directly, for example, a group of patients, in an effort to understand all patients, which is the population. There are many reasons why all patients cannot be studied and thus scientists

have little choice but to study a sample. Consider the example: *disease A in a sample of patients is cured by treatment B, the population of those with disease A is the same as the sample of patients; therefore, the population of patients with disease A will be cured with treatment B.* The syllogism is extended to *my patient is the same as the population of patients with disease A and therefore my patient will be cured with treatment B.* However, it is clear there are many examples where *my patient* was not *cured with treatment B.* Indeed, this would be an example of irreproducibility in the broad sense, and one needs only to look to the drugs, biologics, and devices recalled by the FDA to see this is true. Something must be amiss. The majority of pivotal phase 3 trials that initially garnered FDA approval and later abandoned were not likely to have been type I errors based on fraud; lack of transparency; failure to report reagents, materials, or methods; or statistical flaws within the studies.

There are a number of other fallacies to which the scientific enterprise is heir to. These include the Fallacy of Pseudotransitivity that affects the formulation of hypotheses critical to the Scientific Method, the Fallacy of Affirming a Disjunctive, the Fallacy of Limited Alternatives, and the Gambler's Fallacy. Each will be explored in detail. However, it is very important to note that just because an experiment may have committed a fallacy, it does not mean that the results of the experiment are false, but only one cannot be certain. In that uncertainty lies the risk of irreproducibility.

Scientists should not take umbrage when a risk for a logical fallacy is demonstrated or, more generally, when it is pointed out that the Scientific Method and the scientific enterprise are prone to fallacies. It only means that one cannot have absolute confidence in the result and in that doubt, there is the risk of irreproducibility. Indeed, circumstances exist where a greater risk means a greater chance of new knowledge. Thus, the optimal science may well require a judicious use of logical fallacies, as will be explained in this book.

It should not come as a surprise that the scientific enterprise may be liable to fallacies. Gaining new knowledge is difficult and complicated, particularly in biomedical research. Indeed, it will be demonstrated that many fallacies are critical to the advancement of science, as little progress would be made without them. This may seem counterintuitive but patience in coming to understand the utility of fallacies will be rewarded. The Scientific Method developed for a reason. Statistics evolved for a reason. Both are responses to the fundamental uncertainties of gaining knowledge—any knowledge, but particularly scientific knowledge. Thus, misuse of the Scientific Method or statistics, in many cases, may be symptoms of failures in understanding the challenges of gaining new knowledge—epistemology—that led to the Scientific Method and statistics in the first place. It is not the fault of the Scientific

Method or statistics if its practitioners fail to understand the fundamentals that lead to practitioners misusing the tools. One cannot blame the hammer if a carpenter chooses to use it to saw a piece of wood.

There is a reason why philosophers avoided logical fallacies—they risk irreproducibility. Indeed, one of the most powerful philosophical methods of analysis is to demonstrate that the consequence of an argument results in a contradiction or absurdity, which could be considered an example of irreproducibility in philosophical analyses. Biomedical research experimentation, being a form of argumentation, thus is inherently at risk for irreproducibility by virtue of its necessary trading in logical fallacies. Indeed, the experiment cannot be immunized by statistics against the effects of the inherent logical fallacies. As will be demonstrated in this book, statistics is derivable conceptually from an extension of syllogistic deduction, a logical form, to the partial syllogism—an invalid (not guaranteeing certainty), but useful (creating the possibility for new knowledge), form of logical reasoning.

This book focuses on the epistemic, particularly, logical foundations for biomedical research reproducibility. It is not a primer on statistical analyses. Rather, it examines the implications of the necessary judicious use of logical fallacies on biomedical research experimental designs, including statistical design. The implications are many. Cases are presented that demonstrate a proper use of logical fallacies that decrease the risks of irreproducibilities, yet judicious irreproducibility, such as in a negative study, can provide a contribution that is certain—an example of productive irreproducibility.

The potential for an irreproducible experiment contributing positively to biomedical research is not only helpful, it is fundamental to the advancement of biomedical research. Indeed, productive irreproducibility falsifies the hypotheses, which conveys the greatest, perhaps only, certainty. However, from the reactions voiced in editorials and policy statements, it would appear that irreproducibility is an anathema, a plague that must be avoided at all costs. When "vaccinations" (education) do not work, "quarantine" (rejection of papers and grants) is required. But is this really the case? At least some form of irreproducibility is critical to the success of biomedical research. Perhaps what is critically needed and should be supported is productive irreproducibility with only unproductive irreproducibility avoided. Indeed, understanding what constitutes productive and unproductive irreproducibility is a central theme of this book. In contrast, cases are presented that demonstrate irreproducibilities resulting from poor logic in the experimental design, even if statistically sound, that produce results that are *indeterminant* and of little or no use—an unproductive irreproducibility. Approaches used to mitigate the increased risk of unproductive irreproducibility from the necessary use of logical fallacies are presented.

The epistemic status of statistics is necessarily self-referential and hence internal. Indeed, statistical claims, in themselves, are not ontological (a claim as to reality). Key assumptions are required to be met to maintain the epistemic safety of the logical fallacies that are the basis for the statistical method used. Chief among those assumptions is the Large Numbers Theorem, which holds that with increasing sample size, the measure of the Central Tendency, as measured by the mean (average), approaches a constant (unwavering) number—itself a form of reproducibility. In other words, the Central Tendency, such as the average or mean, becomes stable as the number of measurements included in the calculation of the Central Tendency increases. The convergence onto a stable measure of Central Tendency is thought to provide at least some credibility to the notion that the mean has some status in reality (ontological status). Another important concept is ergodicity, meaning the degree to which measures of the Central Tendency of samples are the same as the Central Tendency of the entire population. For example, if one were to measure the blood pressure from several groups of patients thought to have the same disorder, one would have little confidence if the Central Tendency, for example, the average blood pressure, was different among all the groups. At the very least, the question arises whether one or more groups of patients did not actually have the disorder. Alternatively, it just may mean that the population of all patients with the disorder is not distributed uniformly so that selecting different samples results in very different measures of Central Tendency. If the fish in a lake were distributed evenly (randomly), it would not matter where in the lake one fished. Violations of these assumptions have been demonstrated in the mathematics and physics of Chaotic and Complex Systems, and it is probable that these key assumptions, the Large Numbers Theorem and ergodicity, are not operable in biological Complex or Chaotic systems.

There is increasing evidence that virtually any living biological organism operates in the domain of Complexity and Chaos, in part due to the fact that living biological systems operate far from thermodynamic equilibrium. Typically, an equilibrium is achieved when opposing forces balance each other. In such Chaotic or Complex systems, ergodicity may not apply and thus an application of standard statistics may not be appropriate. Their use nonetheless increases the risk of appearance of irreproducibility, regardless of whether the experiments actually are irreproducible. New statistical approaches are needed. A sketch of what the new statistics may look like is provided.

None of the concerns expressed here should be construed as devaluing biomedical research. This author has been a biomedical researcher and physician long enough to have seen changes in medical treatments that would qualify as miraculous by virtually any standard. This author

has great optimism that the pace of new miracles will only increase in the future. Further, the large majority of other biomedical researchers encountered by this author are dedicated to advancing biomedical research and are honest and honorable. But biomedical research is difficult and challenging at nearly every level. It is easy to trip and fall, particularly over a stone that is not seen, such as an injudicious use of a necessary logical fallacy. It is hoped that this book will help illuminate the path, at least as it concerns the obstacles to the judicious use of logical fallacies.

Thanks to Melissa Revell, whose editing spared the reader the gibberish of my dyslexia. Also, many thanks to Mr. Fred Haer, President of the Greenville Neuromodulation Center, for the continuing support of the scholarship this book required.

**Erwin B. Montgomery, Jr., MD**