CHAPTER

1

Reproducibility in Biomedical Research

Kevin Mullane, Michael J. Curtis, Michael Williams

OUTLINE

Research in the Biomedical Sciences. http://dx.doi.org/10.1016/B978-0-12-804725-5.00001-X

1.1 INTRODUCTION

The remarkable progress and benefits scientific research has brought to human well-being are built on a foundation created by previous researchers. The robustness and consequent value of this edifice is predicated on the notion that researchers adhere to sound practices—the *responsible conduct of research* (NASEM, 2017)—and communicate openly their complete and accurate findings and the methods used to derive them. It is a cumulative process, dependent on both successes and failures, which builds a reliable body to advance scientific knowledge. The process includes the initial discovery, its reproduction by others, and then extension of the initial finding, and where appropriate, correction. Discoveries initially considered as significant or even important might ultimately be tempered or revised as new and improved techniques, tools, and analytical methods are applied that lead to modifications to the initial interpretations. This is a normal part of the research process, and has given rise to the concept that research is validated through the process of scientific self-correction, where "self" refers to the research community at large. That is, independent reproduction of scientific findings provides validity, whereas poorly conceived or badly executed studies will not be reproduced and will be rejected.

The present monograph focuses on the current, highly visible debate on what is viewed as a systematic lack of reproducibility in biomedical research (Economist, 2013a), what it is, its scope, and perceived cause(s), why it matters, and the approaches that are being used to deal with it more effectively. While frequently described as a crisis, Gorski (2016) has taken issue with this terminology noting that "Reproducibility in science is … a chronic problem…. not a crisis," a viewpoint shared by McClain (2013) and many others. Indeed, concerns have been expressed ever since the advent of the first scientific journals in the 17th century, when the eminent scientist philosopher Robert Boyle noted "you will find…many of the Experiments publish'd by Authors, or related to you by the persons you converse with, false or unsuccessful" (Bishop, 2015).

The importance of the reproducibility problem is reflected in the adages, "the demarcation between science and nonscience" (Braude, 1979), "Science isn't science if it isn't reproducible" (Roth and Cox, 2015), "reproducibility, a bedrock principle in the conduct and validation of experimental science" (Casadevall and Fang, 2010), the "coin of the scientific realm" (Loscalzo, 2012), and "If a scientific finding cannot be independently verified, then

it cannot be regarded as an empirical fact. And if a literature contains illusory evidence rather than real findings, the efficiency of the scientific process can be compromised" (Bollen et al., 2015).

The self-correction process of independently reproducing research findings to validate them is an honor system, "built upon a foundation of trust and verification, trust among scientists, who rely on each other's data and conclusions, and trust between scientists and the public, which funds much of the science" (Alberts et al., 2015; Kraus, 2014). The continued effectiveness of self-correction has been questioned as the biomedical research enterprise grows, leading some to argue for a more systematic and transparent approach (Estes, 2012; Morrison, 2014; Pattinson, 2012), since many negative studies are not published, and irreproducible results often remain in the scientific canon (Begley, 2017a).

Given that society has come to depend on the application of findings from basic biomedical research to understand the causes of human diseases, to identify new drug targets, and develop therapeutics to treat disease—not necessarily in that order—if experimental findings cannot be repeated, research becomes a veritable house of cards with a diminution in both its value and its impact on the primary goal of improving human health. If that goal is compromised, and the bond of trust between biomedical research and the public is eroded, it can have deleterious consequences for the continuation of research support (Alberts et al., 2014; Begley and Ioannidis, 2015).

This by no means implies that irreproducible findings should automatically be viewed as fraudulent. As noted by Flier (2017), "Although scientists should and most often do seek to publish reliable results, to expect a standard of certainty before publication, and/or to excessively stigmatize or penalize claims later found honestly to be in error, would diminish progress by replacing a spirit of scientific excitement and daring with professional fear of error…. [requiring researchers to be] ….appropriately tolerant of tentative conclusions and honest errors, while continuously seeking to reduce the latter."

A failure to reproduce an experimental finding can result from a variety of factors, including:

- honest errors;
- the adoption of "detrimental" or "questionable" research practices, colloquially referred to as "sloppy science" that are often enabled by "perverse incentives";
- the introduction of biases, conscious, or unconscious; and
- unknown variables—frequently described as "noise" (Kass et al., 2016)—that by definition cannot be controlled, but include limitations in the tools and techniques for addressing a research question, the skills of the researcher, and multiple undefined factors that can emerge and influence scientific outcomes.

Many of these factors are touched on in this chapter and developed in more detail in subsequent chapters of this monograph.

1.2 DEFINING REPRODUCIBILITY

An unambiguous understanding of precisely what constitutes a problem is essential in its solution. In biomedical research, the term *reproducibility* is used interchangeably with that of *replicability*, often in the same sentence despite the fact that there are subtle but important

differences in their meaning (Baker, 2016a). In fact, there are so many definitions for the activities related to reproducibility that the reader can end up being totally confused as to which term means what in different research disciplines and when/how they should be used (Baker, 2016a).

It is also important to note that much of the discussion on improving reproducibility is informed by experiences in the clinical and social/psychological sciences that may not appear to have immediate relevance to other areas of biomedical research. With respect to clinical studies, due to the inherent responsibility and accountability in testing new drug candidates in human subjects, the research infrastructure is, by design, infinitely more complex, transparent, and multidisciplinary as well as better funded than preclinical research. The randomized controlled trial (RCT) format of clinical trials represents the gold standard in biomedical research (Bothwell et al., 2016) and has been proposed as a model for animal experimentation (Henderson et al., 2013; Muhlhausler et al., 2013) while being used as the basis to develop the *ARRIVE* and related animal research guidelines (Chapter 5.10). However, the resources required to implement RCT-like standards in preclinical research will be a major challenge; one obvious source may be the estimated $28 billion that is wasted annually in the United States on irreproducible research due to poor design, execution, and reporting (Freedman et al., 2015). In the field of social psychology, reproducibility initiatives were, until recently, rare (Earp and Trafimow, 2016; Schmidt, 2009), the result of both a perceived "lack of prestige" in "copying" the work of another researcher, or a lack of respect for the competence/eminence/prestige of the scientist who had conducted the original study.

Data that are *reproducible* represent those that are *similar* to one another and reflect "the extent to which consistent results are obtained when produced repeatedly" (Casadevall and Fang, 2010) with Fang further noting that these results need to be "robust enough to survive various sorts of analysis" (Baker, 2016a). Data that are *replicable* are *exact* copies of one another. However, in biomedical research only reproducible data has value since replication, a term used in the field of biomarkers and clinical chemistry is, in practical terms, unfeasible as few, if any, experiments can actually be replicated due to biological variability attributed to known and unknown differences in experimentation (Nosek and Errington, 2017). Adding to this confusion in terminologies, in the biomarker field an additional term, r*epeatability,* is used that "refers to the outcome of measurements that are performed under *the same conditions*, while.. [using]… reproducibility to… [describe]… the outcome of measurements performed *under different conditions*" (Chau et al., 2008).

In their seminal editorial on reproducibility, Casadevall and Fang (2010) noted, based on definitions used in the field of information technology (Drummond, 2009), that "reproducibility refers to a phenomenon that can be predicted to recur even when experimental conditions may vary to some degree. On the other hand, replicability describes the ability to obtain an identical result when an experiment is performed under precisely identical conditions." Another term that is often used instead of reproducibility, is that of *generalizability* "the persistence of an effect in settings different from and outside of an experimental framework" (Goodman et al., 2016). Thus the terms "reproducible" and "generalizable" also speak to the *robustness* of a finding—that the same qualitative outcome is observed under different conditions to give confidence that the effect is real, and not an artifact consequent to some particular feature of a test system.

The reader may wonder about the relevance of these apparently semantic distinctions for addressing the issue of irreproducibility and why it matters. The answer lies in the expectations from research findings. If the expectation is that reconstructing a published experiment will yield quantitatively *exactly* the same results, then the experimenter will likely be disappointed. Indeed, Nosek and Errington (2017) have noted "There is no such thing as exact replication because there are always differences between the original study and the replication.... As a consequence, repeating the methodology does not mean an exact replication, but rather the repetition of what is presumed to matter for obtaining the original result."

If the broad research finding can be repeated independently to reach the same conclusions, then it bodes well that the effect is real and could be extended into additional research, and possibly clinical, settings. In this regard, Begley and Ioannidis (2015) also concluded that while "There is no clear consensus as to what constitutes a reproducible study….. it seems completely reasonable that the one or two big ideas or major conclusions that emerge from a scientific report should be validated and withstand close interrogation."

Clearly an area exists between adequately adopting key elements of the original study to ensure reproducibility and determining the validity and robustness of the findings by exploring alternative settings. Thus the intention is to reproduce in a closely matched setting and then extend the observations using alternative approaches/techniques that are still appropriate to meet the goal of a study (Flier, 2017).

Attempts to reproduce a finding often tend to overlook key elements of the original study and have proven to be a topic of debate regarding the Open Science Collaboration (OSC) report, "Estimating the reproducibility of psychological science" (OSC, 2015) (Chapter 5.13.2.3). One major issue has been the use of a different subject population in the reproducibility studies from that used in the original study (Gilbert et al., 2016). For instance, the replication/reproducibility attempt for an original study that asked Israelis to imagine the consequences of military service was replicated by asking Americans to imagine the consequences of a honeymoon, which, perhaps unsurprisingly, failed (Mathur, 2016). Reproducible *findings* represent the goal, while the relevance of a *study* is, by definition, limited until the findings can be independently and robustly reproduced in other different, populations. Replication, once again can be considered to have limited intrinsic value in biomedical research.

1.2.1 A New Lexicon for Reproducibility

As a consequence of the extensive semantic confusion in the use of the terms *reproducibility* and *replicability*, with *repeatability* being used to cover all eventualities, Goodman et al. (2016) have proposed a "new lexicon" to better describe issues in reproducibility using qualifiers for methods, results, and inferential reproducibility.

1.2.1.1 Methods Reproducibility

Methods reproducibility was defined as the reproduction of experimental procedures including design, execution and data analysis in reports of an original study. For example, it is often assumed that the purchase of a product (e.g., a cell-line or antibody) from a catalog implies that the vendor has performed some quality control on the product, but they are often just a repository and distribution center for multiple products from different sources and make no commitments to the accuracy or integrity of the materials provided. It is up to

the investigator to ensure the validity of any reagent used (Chapter 2.5.2). Moreover, cell-lines drift, become contaminated, undergo genomic and epigenomic alterations, etc., all of which can modify experimental results without careful attention to their fidelity. This level of detail is frequently missing from publications (Vasilevsky et al., 2013) often for the simple reason that it has not been done (Freedman and Gibson, 2015; Freedman and Inglese, 2014; Nardone, 2008; Neimark, 2014, 2015).

1.2.1.2 Results Reproducibility

Results reproducibility was defined as the situation where an experiment is conducted to confirm a previously reported finding, the procedures used for which are as closely matched to those used in the original experiment as is possible. However, when a study is not reproduced, frequently the original study is found to be underpowered and/or provides a low signal-to-noise ratio and is incapable of discerning a true effect, such that the criteria for defining results that are "the same" need to be so loose that they lose meaning. Instead, "the paradigm for accumulating evidence…[is] more appropriate than any binary criteria for successful or unsuccessful replication," a point often overlooked by proponents of formalized reproducibility initiatives (Chapter 5.12). If the effects reported in a study are weak and group sizes are inadequate then there is a high probability that a finding will be false, regardless of whether it reproduces data from a previous finding.

Another important aspect is the overreliance on, and limited understanding of, what precisely a p value denotes. While this issue is developed more fully in Chapter 3, it should be noted that the probability of repeating a study that is just significant at the 5% level (i.e., $p = 0.05$) is only 50% due to random variation of the p value, but does not necessarily undermine the credibility of the first experiment (Button et al., 2013; Goodman et al., 2016; Horton, 2015).

1.2.1.3 Inferential Reproducibility

Inferential reproducibility was described by Goodman et al. (2016) as the "most important" of the three types of reproducibility since it "refers to the drawing of qualitatively similar conclusions for either an independent replication … (i.e., reproduction)… of a study or a reanalysis of the original study," the latter assuming that the information from the original study is available which is infrequently the case (Engber, 2016). The authors coined the term inferential reproducibility "because scientists might draw the same conclusions from different sets of studies and data or could draw different conclusions from the same original data… even if they agree on the analytical results." The issue of drawing different conclusions from the same data serves to illustrate that this new terminology may take time to be accepted into mainstream biomedical research especially as it invokes the use of Bayesian paradigms, an area of statistics where many researchers are either uncomfortable or dismissive (Colquhoun, 2014).

1.2.1.4 Bayesian Paradigms in Reproducibility

Goodman et al. (2016) concluded their proposal for a new lexicon in scientific reproducibility on the topic of "operationalizing truth" noting that "replication" per se is not the objective of reproducing an experiment but rather "whether scientific claims based on scientific results are true since if a finding can be reliably repeated, it is likely to be true." Unfortunately, terms such as "true" and "false" can provoke an emotive connotation that impugns the integrity of the researcher while it is only addressing the reproducibility of an outcome.

As the traditional frequentist approach to statistics (Chapter 3.6) cannot assign any probability to the truth of a finding, Goodman et al., advocate the use of Bayesian statistics where "The probability that a claim is true after an experiment is a function of the strength of the new experimental evidence combined with how likely it was to be true before the experiment," an important concept that is often overlooked in the reproducibility debate. Thus, rather than relying on a single reproduction of a research finding (or not), the current *modus operandi* of the various formal *Reproducibility Initiatives* (Baker and Dolgin, 2017) (Chapter 5.12) reflect the hypercompetitive aspects of 21st century biomedical research (Alberts et al., 2014), where the deference given to the first published, peer reviewed-finding (its "priority") leads to its reproduction being viewed as a test of the original published finding rather than a part of the continued testing of the original hypothesis via the process of self-correction (Alberts et al., 2015). In this context, Goodman et al. note that "the aim of repeated experimentation is to increase the amount of evidence, measured on a continuous scale, either for or against the original claim ...[resulting in]…strong *cumulative* evidence with each independent data set adding to a final outcome that can confirm or refute a hypothesis." Such a strategy was espoused in 1620 by Francis Bacon, considered the grandfather of the empirical scientific method. Based on performing experiments, making observations, and analyzing the outcomes, Bacon observed that "Now my method, though hard to practice, is easy to explain, and it is this. I propose to establish progressive stages of certainty" (http://www.constitution.org/bacon/nov_org.htm).

1.3 DISCIPLINE SPECIFIC TERMINOLOGY IN THE BIOMEDICAL SCIENCES?

As previously discussed, there is considerable confusion in the definitions used when discussing reproducibility (Baker, 2016a), much of it discipline specific. In 2015, a National Science Foundation (NSF) workshop on the topic of enhancing the reliability of research in the social and behavioral sciences (Bollen et al., 2015) defined *reproducibility* as "the ability of a researcher to duplicate the results of a prior study using the same materials and experimental design and analysis as were used by the original investigator"; an additional condition of this definition was the reuse of the same raw data set that was used to generate the original finding. *Replicability* was defined as "the ability of a researcher to duplicate the results of a prior study if the same procedures are followed *but new data are collected"* (author emphasis added). Of additional note, both definitions include the term "duplicate" adding to the confusion since the latter is defined as "exactly like something else" such that these definitions of reproducibility and replicability appear to differ on only a single point—whether new data is generated. This appears to be a situation unique to the social, behavioral, and economic sciences—the subject of the NSF workshop—and may be reflective of the behavioral sciences having lower methodological consensus and higher noise (Fanelli and Ioannidis, 2013). Interestingly, the NSF recommendations were also "opposed" (or ignored) in the OSC paper on reproducibility in the psychological sciences where *new studies* that attempted to repeat original studies were described in terms of the NSF definition of *reproducibility* rather than *replicability* (OSC, 2015).

Finally, in response to issues with reproducibility failures in the initial reports from the *Reproducibility Project: Cancer Biology* initiative (Baker and Dolgin, 2017), a successor to the

OSC study on reproducibility in the psychological sciences (OSC, 2015), Nosek and Errington (2017) qualified their replication (e.g., reproducibility) process with the terms direct and conceptual. *Direct replication* was defined as "attempting to reproduce a previously observed result with a procedure that provides no a priori reason to expect a different outcome [using] protocols from the original study [that] are followed with different samples of the same or similar materials: as such, a direct replication reflects the current beliefs about what is needed to produce a finding"; the definition of *Conceptual replication* concerned "a different methodology (such as a different experimental technique or a different model of a disease) to test the same hypothesis: as such, by employing multiple methodologies conceptual replications can provide evidence that enables researchers to converge on an explanation for a finding that is not dependent on any one methodology."

In summary, reproducibility and replicability, as well as repeatability, are terms that mean the same thing to many researchers and are used interchangeably in the literature. Since there is a difference between the exact copying of a study, and the confirmation of a finding, there have been many attempts, mostly unsuccessful, to imbue a clear distinction between the two and the implications this has in real world experimentation (Nosek and Errington, 2017). Since reproducibility is what matters in assessing the broader truth of a research finding (Goodman et al., 2016), this is the descriptor that will be used by the authors to describe the independent confirmation of a research finding—its self-correction; however, the terms replication and replicability will inevitably appear in citations from the literature, with the majority of this usage being the result of common usage rather than scientific accuracy.

1.4 EXPERIMENTAL FACTORS IN ADDITION TO STATISTICS THAT AFFECT REPRODUCIBILITY

The ongoing debate on the misuse of statistical analysis as a key—or more often *the* key—factor in the reproducibility debate (Colquhoun, 2014; Kass et al., 2016; Marino, 2014; Motulsky, 2014) has tended to overshadow other causal factors some of which are binary and have a greater impact on research outcomes than an inadequate appreciation of statistical methodology. This includes (Chapter 2) flawed experimental design (Curtis et al., 2015; Glass, 2014; Holder and Marino, 2017; Ruxton and Colegrave, 2011) that includes the lack of a credible hypothesis, inadequate powering (Marino, 2014; Motulsky, 2014); the absence of appropriate controls, both positive and negative (Begley, 2013; Begley and Ioannidis, 2015); the use of unvalidated antibodies (Baker, 2015; Bradbury and Pluckthun, 2015) and inappropriate/nonselective concentrations/doses of reference tool compounds (Arrowsmith et al., 2015); an absence of cell line authentication (Almeida et al., 2016; Geraghty et al., 2014; Nardone, 2008; Neimark, 2014, 2015), lack of blinding and randomization (Hirst et al., 2014; Ruxton and Colegrave, 2011; Suresh, 2011), and an overinterpretation/overgeneralization of the results obtained from animal studies (Section 1.11.1). These are what are referred to as "detrimental research practices" or simply as "sloppy science."

As noted not all irreproducible studies are due to fraud, bias, or detrimental research practices and that resolution of these issues will minimize all questions regarding reproducibility.

1.4.1 Unknown Unknowns Affecting Reproducibility

Despite best efforts and taking extraordinary measures to limit potential sources of variability, attempts to replicate studies can still lead to statistically significant differences in outcomes, for reasons that have yet to be identified, and are referred to as unknown unknowns.

1.4.1.1 The Crabbe Mouse Study

In an examination of the behavioral phenotypes of several genetically inbred mouse strains and one null mutant at three different laboratories in Albany, New York, Edmonton, Alberta and Portland, Oregon, Crabbe et al. (1999) took steps to control for; animal strains (by using the same supplier); test apparatus; testing protocols; handling and cage environment (including materials, bedding; incandescent light exposure; number of littermates; cage changing frequency) etc., with behaviors being assessed simultaneously in the same sequence at the same time of day at all three sites. Performance of the mice was assessed in six validated behaviors—locomotor activity, anxiety in the elevated plus maze (EPM), rotarod performance, swim test, cocaine-induced locomotor activation and ethanol preference, and powered to a 90% level. The expectation, given the extensive planning involved, was that the data generated by each laboratory would be similar, but this was not what was observed.

In the cocaine-induced locomotor activation test, mice treated with cocaine in Portland moved 600 cm more than control animals did. In Albany, cocaine-treated mice moved 701 cm, not markedly different from the results obtained in Portland. However, those tested in Edmonton moved 5,000 cm (nearly an order of magnitude difference that did not require any estimate of standard deviation or statistical analysis to provoke consternation). Other digressions were observed in the EPM, leading the authors to conclude that the results were idiosyncratic to a particular laboratory (Crabbe et al., 1999).

The findings of this study resulted in many behavioral scientists concluding that the standardization of experimental conditions was an exercise in futility (Van der Staay and Steckler, 2002). However, in the 17 years since these studies were conducted, previously unknown variables that may have contributed to the experimental differences have come to light. These include the influence of pheromones produced by researchers on animal behavior (Sorge et al., 2014), the animal microbiome (Bahrndorff et al., 2016; Cryan and O'Mahoney, 2011; Dinan et al., 2015; Ezenwa et al., 2012) which differs in its composition, and effects on host phenotype based on environmental factors (Rogers et al., 2014), background passenger mutations in genetically modified mice (Vanden Berghe et al., 2015) as well as more mundane causes including circadian rhythms (Drucker, 2016). These factors, all of which have been reported to alter animal phenotypes, especially behavior, may potentially explain the discrepancies in the results seen in the Crabbe et al. study, while there may still be other unknowns yet to be identified.

1.4.1.2 Reproducibility Confounds in RNA Interference

While it can be argued that in vivo experiments in general, and behavioral studies in particular, can be subjected to marked variations in outcomes that muddy interpretations due to the large number of uncontrolled and uncontrollable variables, in vitro and even molecular biology studies are not immune to similar vagaries. Two identically designed RNA interference-based whole genome screens to identify host factors that support yellow fever

virus propagation in human cells using high-content cell-based imaging conducted 5 months apart by the same investigators revealed different hit lists with only approximately 40% overlap (Barrows et al., 2010). An additional confounder, the method of analysis also significantly impacted measures of intra- and interassay reproducibility despite the four analytical tools utilized being accepted and used routinely in the field. Reasons for the low reproducibility of hits identified from the two studies remain speculative, while the analytical inconsistencies are an example of a known unknown, where the issues surrounding the analytical methods are known but the information necessary to identify the "correct answer" is not.

One problematic issue of such RNA interference assays is "false discovery rates"; false negatives that miss valid hits, and, in particular, false positives that erroneously assign activities that are not real. A second-generation RNA interference library computationally optimized to decrease the incidence of false discoveries was compared to results previously obtained by the same research team in 2005 using the same cell line, reporters, and experimental design, but that had used a first-generation library to study the *Drosophila* JAK/STAT pathway (Fisher et al., 2012). Since analytical tools had also improved in the intervening period, the results from 2005 were also subjected to reanalysis. While from the original study, 134 hits were identified, there were 42 in the follow-on study, but only 12 targets were common to both screens. While this can partially be explained by improved library design to reduce off-target effects, the second-generation library still showed 31% false positive hit rates upon rescreening, and other unknown factors likely contributed to the variability. The study highlights the difficulties in reproducing such studies, particularly as older libraries become superseded by newer ones and are no longer available. As the authors (Fisher et al., 2012) concluded, "even the most sophisticated screening approaches are only a tool to identify genes that potentially interact with a chosen assay system," and these require further validation and confirmation.

1.4.1.3 Caenorhabditis Lifespan

A consortium, the *Caenorhabditis Intervention Testing Program* (Lucanic et al., 2017) was established to assess the impact of genetic backgrounds and compounds on lifespan in different *Caenorhabditis* strains and species rigorously adopted common procedures across three research sites that minimized intersite variability. However, variation in reproducibility at any one site was around 15%, which proved to be of a similar or even greater magnitude to the impact of different genetic backgrounds on the longevity of 12% among species and 8% among strains within species, that might obfuscate their influence or that of chemicals intended to prolong lifespan. The investigators attribute this intralaboratory variation "to unidentified and apparently subtle differences in the assay environment, which vary similarly within each laboratory."

Few investigators are willing to conduct confirmatory studies using new tools or methods that might repudiate their earlier work or conclusions, and which could result in findings that cannot be explained, are difficult to publish, and question the direction and fidelity of their research efforts. Consequently, the full impact of unanticipated, unidentified, and unknown variables on the reproducibility problem cannot be quantified. However, the few examples described here show that problems of reproducibility are not due only to bias or poor experimental techniques or design, but there is a wide range of mitigating factors. While these must be distinguished from issues of "sloppy science," they are important regarding the robustness

of scientific discoveries and their translation to the clinic. Only the most robust findings that transcend the minutiae of animal husbandry or incubation conditions or the like stand a chance of going on to benefit patients.

1.4.2 Known Unknowns: Tacit Expertize

As in other domains of human performance, whether it is knitting a scarf, *haute* cuisine *a la* Bocuse or Robuchon, crafting a musical instrument (think Stradivarius), tuning a piano or performing complex surgery, some biomedical researchers are better than others in the routine execution of experiments. Such individuals reliably derive outcomes that others cannot, a phenomenon described as "superior technical skills" (Stroebe et al., 2012). This applies to animal surgery in the areas of pain, stroke, tumor xenografts, etc. and in some laboratories, this has led to a designated researcher being tasked with conducting a particular procedure. While this may suggest that there are irregularities in experimental procedures, in such instances, independent oversight has repeatedly determined that this is not the case. Rather, the researcher as a result of their experience, continued practice, and intrinsic mental and physical coordination abilities has developed skill sets that others lack and are not easily taught—a gift of nature as it were. Thus, in the field of biomedical research, some laboratories are well known for their unique ability to produce certain types of data and equally widely known for not manipulating their results. As a result, their techniques and the data they produce become the benchmark for all others working in their field even though in a strictly rational sense, their findings can be described as irreproducible.

1.4.3 Diminishing Effects: Regression to the Mean

As studies progress and additional information is provided it is not unusual for previously reported or observed activities (e.g., of a therapeutic agent) to have smaller effects. This phenomenon, first noted by Francis Galton in 1886, describes the tendency for data-points that are outliers to move toward a population mean over the course of repeated measurements—an event termed "regression to the mean" (Sen, 2011). Serikawa (2015) has made the analogy to baseball players who, during the season, can go through "hot" or "cold" streaks in batting averages or on-base percentages, while the full season is the better indicator of true performance.

In biomedical research, this trend is particularly evident in clinical trials of new drug candidates since the number of participants in clinical studies escalates as the therapeutic progresses, so the increased number of data-points reveal the true effect of the drug, and outcomes might differ markedly from the earlier, smaller Phase II trials, and clearly have a major impact on their reproducibility.

Several explanations have been proposed to rationalize the diminishing effects observed in randomized clinical trials. The first, as indicated, simply relates to sample size, and this is the key factor in explaining the decline effect or "regression to the mean." Ioannidis (2006) calculated that a median sample size of 80 patients (the average Phase IIa study), and prestudy odds of 1:10, using $\alpha = 0.05$ with 20% power, the outcome has only a 28% likelihood of being true. Consequently, the chances of repeating the outcome in a second study, or a larger Phase IIb or Phase III trial, are slim. Moreover, there is also a "clinical trial effect" where

patients are monitored continuously and adherence to treatment is fostered, that can differ from "real world" situations in a more heterogeneous group of patients (Menezes et al., 2011) (Section 1.11.2.2.1).

However, diminishing effects in clinical trials, and hence poor reproducibility, can also be attributed to other influences. One such factor is *patient selection*, based on entry criteria for a specific study and random (biological) variation. Many chronic disorders, such as arthritis, asthma, hypertension, depression and back pain, are not static, but fluctuate in terms of their severity. Clinical trials recruit patients based on specific entry criteria, and often patients seek out new treatments when their disease has flared up and they are not getting adequate relief from their standard medications. However, during the course of the study, the natural variation might cause their symptoms to recover from that peak, which cannot be differentiated from the effectiveness of the treatment.

This is exemplified in Fig. 1.1 (McGorry et al., 2000) as discussed elsewhere (http://www.dcscience.net/2015/12/11/placebo-effects-are-weak-regression-to-the-mean-is-the-main-reason-ineffective-treatments-appear-to-work/). These researchers followed the natural history of patients with low back pain over 5 months, recording their daily pain perception on a 10-point scale. Despite having free access to pain medication as needed, the patients show marked fluctuations in their symptoms, and this background variability or "biological noise" makes it extremely difficult to discern the effect of a new treatment, or obtain reproducible data.

Some, but not all, clinical trials have a "run in" period to stabilize existing treatment and ensure patients still meet the entry criteria. However, for practical considerations the "run in" period is often short and not based on any detailed knowledge or consideration of the dynamics of the individual's disease, while there is also pressure to maintain recruitment even if participants now fall a little outside the criteria.

Clearly in the example provided in Fig. 1.1, outcomes and conclusions are going to vary depending on whether or not patients were recruited or drug effects measured at a peak, trough, or midpoint. Indeed, a metaanalysis of 118 clinical trials related to treatments for nonspecific lower back pain reported a similar modest improvement in pain scores over a few weeks, regardless of treatment type or even if any treatment was involved (Artus et al., 2010). A corollary is that posthoc analyses of clinical trial data often interpret this biological fluctuation as indicative of a patient "subgroup" that derives significant benefit, when it has nothing to do with the treatment (http://www.dcscience.net/2015/12/11/placebo-effects-are-weak-regression-to-the-mean-is-the-main-reason-ineffective-treatments-appear-to-work/).

Biological noise is supposedly offset by inclusion of a placebo-control group as a statistical comparator. However, small sample sizes can also skew the placebo data, while patient selection according to specific entry criteria can bias the dataset and contribute to misleading conclusions. This problem is exemplified by Sen (2011) modeling hypertension based on diastolic blood pressure readings taken at baseline and again at a later time point, in the absence of any treatment. If the data were cut according to a baseline reading for hypertension greater than 95 mm Hg, then all that could be measured were subjects who began the study hypertensive and either remained hypertensive or became normotensive. In this group, a drop in blood pressure was observed. But this analysis missed out the patients who started out normotensive and then became hypertensive. Including this group showed there was no change in mean blood pressure. So, the entry criterion of patients needing to be hypertensive to participate in the study introduced a bias that erroneously indicated that a fall in blood pressure occurred.

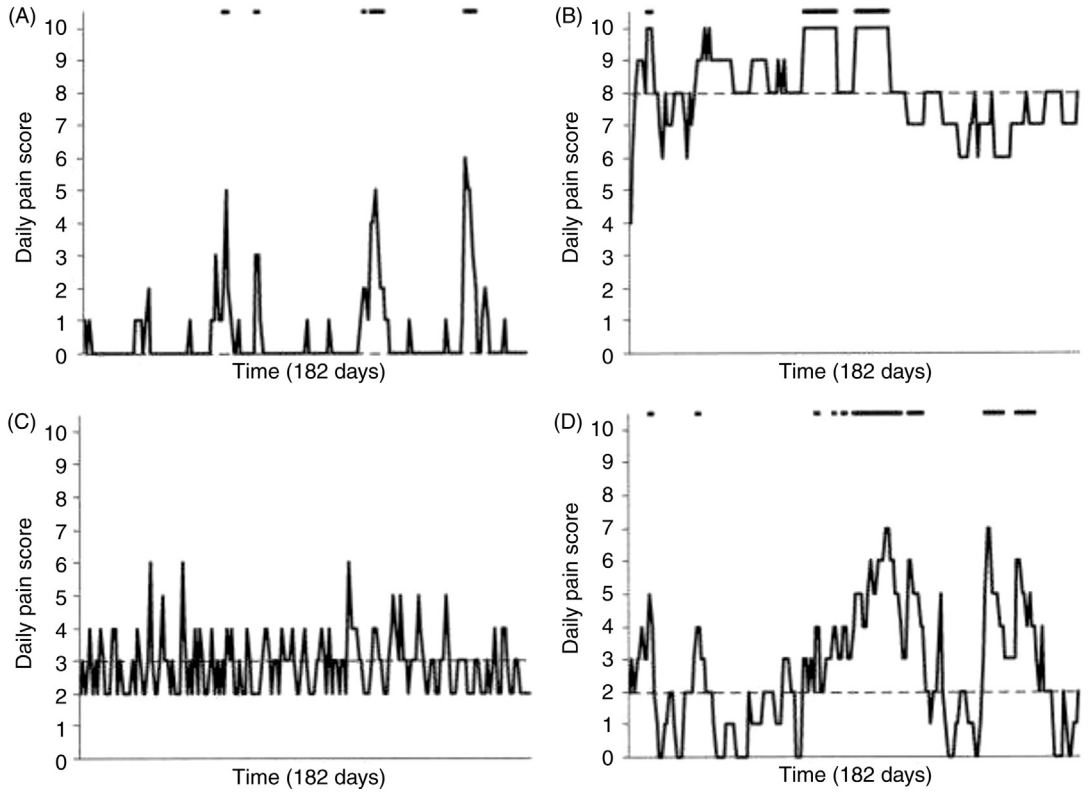

FIGURE 1.1 **Daily recordings of pain intensity, measured on a scale from 0 (no pain) to 10 (as bad as could be), in 4 of the 94 subjects with low back pain (LBP) tracked for 6 months by McGorry et al. (2000).** (A–D) The *dashed bars* at the top of each graph depict the occurrence and duration of episodic "flare-ups," defined as 2–9 consecutive days where the pain score was at least two units above the median score for the 6 months. Despite free access to pain medications as needed, the high degree of variability between subjects and over time is readily apparent, with the authors concluding that LBP is not static and that both its intensity and episodic nature profoundly affect the patient's ability to function. *Source: From McGorry, R.W.,Webster, B.S., Snook, S.H., Hsiang, S.M., 2000. The relation between pain intensity, disability, and the episodic nature of chronic and recurrent low back pain. Spine 25, 834–841, used with permission of Woulters Kluwer.*

Changes observed in patients taking a placebo are often referred to as "the placebo effect," thought to confound many clinical trials and have often been utilized by early stage biotechnology CEOs to justify why their treatment did not work as they predicted. However, a comprehensive review of placebo versus no treatment from 202 clinical trials covering 60 conditions (Hróbjartsson and Gøtzsche, 2010) found that placebo interventions could influence patient-reported outcomes, especially in the areas of pain and nausea, but in general had no important effects. The large placebo effects reported in clinical studies were largely due to issues of experimental design and bias, limited sample sizes, and large outlier effects.

Other explanations for a decline effect over time include changing treatment patterns, since new therapies are provided on a background of the standard of care; while a publication bias of reporting only positive studies might lead to the perception of a diminishing

effect as corresponding negative data necessary for regression to the mean determinations are not available. Funnel plots of the treatment effect charted against precision (e.g., standard error) for individual studies have been proposed to help determine publication bias based on whether they are symmetrical (suggestive of inclusion of negative or neutral studies) or asymmetrical (omission of such studies); although caution must be exercised in their application and interpretation (Lau et al., 2006).

1.4.4 Overinterpretation of Effects on Secondary Endpoints in Studies

Another cause of poor reproducibility between clinical trials relates to overinterpretation of the effects on secondary endpoints. Moyé and Deswal (2002) provide excellent examples where drugs to treat congestive heart failure, including vesnarinone, losartan, amlodipine, and carvedilol showed a benefit in one clinical trial that disappeared or reversed in a second study specifically designed to confirm the first observation. In most of these cases, mortality was used in the first trial as a secondary endpoint, while the study was principally designed and powered around the primary endpoint. The authors pointed out that "Type 1 errors are unacceptably high when subsidiary analysis results are proclaimed as positive when there was no prospective type 1 error statement," and that "the cumulative type 1 error increases with the number of statistical evaluations"—that is, looking at different endpoints; and they reiterate the need to distinguish between exploratory and confirmatory analyses. It is likely that the same confound occurs in preclinical in vivo studies where multiple endpoints are used while the study is designed and powered around a single primary endpoint.

1.5 THE IMPACT OF THE INTERNET ON THE EVOLUTION OF RESEARCH PRACTICES—DATABASES AND THIRD PARTY ANALYSES

The issue of reproducibility is somewhat different when dealing with projects involving "big science," like the NIH Big Data to Knowledge (BD2K) Initiative (https://datascience.nih.gov/bd2k). Large-scale datasets from genomic, epigenomic, transcriptomic, and proteomic analyses (to name but a few) are routinely deposited in publicly accessible databases, as mandated by granting authorities and journals (Chapter 5.3.2.6). Two genomic databases, the *Gene Expression Omnibus* (GEO) at the US National Center for Biotechnology Information, and *ArrayExpress* at the European Bioinformatics Institute, together house the data and links for over 30,000 published studies comprising well over 1,000,000 samples, from more than 13,000 laboratories—and with double digit increases each year (Rung and Brazma, 2013; https://academic.oup.com/nar/article/41/D1/D991/1067995/NCBI-GEO-archive-for-functional-genomics-data-sets). In 2011, the *ArrayExpress* database alone was accessed by approximately 1,000 different users each day (Rung and Brazma, 2013) while another major database, the *Kyoto Encyclopedia of Genes and Genomes* (KEGG), was visited 400,000–500,000 times a month in 2014—a figure that had doubled over the preceding 3 years (http://www.kegg.jp/kegg/docs/plea.html). Similarly, the *Protein Data Bank* and *Uniprot* had nearly 750,000 visitors accessing 20 million pages in just 1 month, back in 2008 (Howe et al., 2008). Another way of measuring the impact of these databases is to look at the number of third party publications

that emanate from the archived material. A *PubMed* search for articles referencing any of the 2,711 datasets deposited in GEO in 2007, led to the estimate that there were more than 1,150 such articles by the end of 2010 (Piwowar et al., 2011). In just 1 year, 2011, 90 publications were identified that cited and used data from any of the five *ArrayExpress* studies published in database issues of the journal *Nucleic Acid Research* (Rung and Brazma, 2013), and it was concluded that the actual number was probably much higher since many studies do not cite the original publication, only the data accession number. The number of third party publications is increasing and it is suggested that one quarter of studies now use publicly available datasets rather than conducting new experiments when addressing biological questions (Rung and Brazma, 2013). Although some pejoratively call such scientists "research parasites" (Longo and Drazen, 2016), and only approximately 3% of metaanalyses are considered well conducted and clinically useful (Ioannidis, 2016a), this is a burgeoning field that has become a practical reality, that brings with it several unique issues surrounding issues in reproducibility.

Interrogating such enormous datasets allows questions to be asked beyond that of any original study with its more limited samples that were designed for a specific, limited query. One common application is to increase analytical power by pooling data from multiple genetic analyses to unmask disease-associated genes and mutations that might not be identified in any single study (Panagiotou et al., 2013; Price et al., 2015; Torrente et al., 2016). Selecting a cancer gene expression dataset of 28,000 samples in *GEO* and *ArrayExpres*s obtained using the same Affymetrix chip led to the identification of 1,285 potential cancer related genes in solid tissue samples, many of which were novel, while gene enrichment analysis identified some cancer pathways that are tissue-specific (Torrente et al., 2016). Gene variants with small effect sizes identified by genome-wide association studies (GWAS) include *HMGCo-A reductase* and *PCSK9* (Proprotein convertase subtilisin/kexin type 9) for LDL cholesterol levels (Global Lipids Genetics Consortium, 2013), and *PPARγ* (Peroxisome proliferator-activated receptor gamma), and *KCNJ11* (Potassium Voltage-Gated Channel Subfamily J Member 11) for Type 2 diabetes (Morris et al., 2012), representing important drug targets that would not be identified by single studies but required the increased power of enlarged datasets.

Entry into the era of predictive, personalized medicine requires information not only on the individual, but how they fit into specific groups. It has been estimated that within the next 10–15 years, the genes of 1 billion people around the world will have been sequenced, and, in many cases coupled with electronically captured medical records (Price et al., 2015). Moreover, smartphone-app enabled health research has just begun (Chapter 6.4.2.6), for example, with curated data shared by Sage Bionetworks from more than 9,000 participants in a study of Parkinson's disease (Wilbanks and Friend, 2016). This is a tool that has the potential to add enormous amounts of patient data into the mix. If issues of data integrity, accessibility, integration, and interrogation can be overcome, this will enable determination of how a particular mutation can influence disease risk and progression, define treatment options and predict outcomes. To realize such a goal requires overcoming several impediments that weaken all aspects of "Big Data" and publicly available databases—namely issues surrounding inputted data quality and integrity; annotation, curation, and accessibility of information within the databases; cross-database integration for a holistic and amalgamated view of a patient; and defined analytical tools that provide accurate and relevant interpretation of the datasets. As articulated by Howe et al. (2008) "Biology today needs more robust, expressive, computable, quantitative, accurate and precise ways to handle data."

1.5.1 Data Input

In the context of archiving experimental data for further use it is particularly important to ensure that it is accurate and complete, since any biases can infest the dataset and be perpetuated in any secondary analyses. As discussed in detail later (Section 1.10.6.5), the data from microarray analyses of gene expression in cancer cell-lines and sensitivity to treatments (Hsu et al., 2007; Potti et al., 2006), was uploaded to GEO, where independent reanalysis of the datasets revealed significant flaws (Baggerly and Coombes, 2009). While this particular example proved to be due to overt fraud, it nonetheless represents an important application of the databases to check for reproducibility concerns. Attempts to reproduce 18 microarray-based studies published in *Nature Genetics* in 2005–2006, with data deposited in *GEO* and *ArrayExpress*, were only successful in 2 cases, while 6 others were "partially" reproduced and 10 not at all, leading to the conclusion that reproduction of such studies is problematic and that stricter rules around the quality and completeness of the data deposited in such databases were required (Ioannidis et al., 2009). Steps have been taken to improve and enforce guidelines for data submission, but there is still need for further improvement (Rung and Brazma, 2013). Moreover, it is apparent that the same control datasets are sometimes used for multiple experiments, so reappear in the database, and can give a distorted perspective. The *Cancer Proteomics Consortium* spent their first 10 years developing standards and ensuring reproducibility before beginning to publish studies, giving an indication of the efforts required.

1.5.2 Data Curation

Many of the databases are not merely repositories, but process, analyze, and annotate the information so that it can be recalled and interrogated effectively. Often the submission of results to the database is performed by the bioinformatician who analyzed the data, rather than the experimentalist who conducted the study, the former of whom might lack detailed knowledge regarding the experiments and protocols. Relying on authors to annotate their own data failed (Howe et al., 2008). To collate all of the information so that access is facile, logical, and user-friendly is an enormous undertaking, well beyond the capabilities of manual curation, so utilizes a range of computational tools that include annotation (Howe et al., 2008), text mining (Singhal et al., 2016), visualization, and querying (Welter et al., 2014), although accuracy remains a challenge. One example of the complications of annotation is the human gene *CDKN2A* (Cyclin Dependent Kinase Inhibitor 2A) that has 10 synonyms in the literature, one of which, *p14*, is also a synonym for 5 other genes (Howe et al., 2008). Limitations in annotation are regarded as one of the key reasons for the poor reproducibility of high-throughput gene expression studies (Rung and Brazma, 2013).

Responsibility for maintaining and updating databases with revised information is frequently unclear. Genetic screening for severe, orphan childhood recessive diseases revealed that a large proportion of literature-annotated disease mutations warehoused in publicly accessible databases, such as the *Human Gene Mutation Database and Online Mendelian Inheritance in Man* were "incorrect, incomplete, or common polymorphisms" (Bell et al., 2011). While 12%–13% of such mutations were simply incorrect, a further 74% could be accounted for by simple substitutions, many of which were erroneously annotated as disease mutations. As noted by the authors, without an accurate and authoritative database, progress

toward prevention, diagnosis and treatment of these diseases will be severely hampered (Bell et al., 2011). The reality is that curators of many databases wrestle with keeping them updated and accurate (e.g., Omenn et al., 2015; Rupp et al., 2016; Weichenberger et al., 2017).

The costs associated with curating and maintaining the multitude of databases are significant, such that after a while many become "stale" despite still being accessible, and it can be challenging to ascertain when they were last updated. Even KEGG, a highly regarded and widely used database, has had to resort to a subscription and licensing model to supplement the grants and maintain viability.

1.5.3 Cross-Database Integration

Most genetic associations with human disease occur in gene regulatory regions rather than the protein-coding regions, so extrapolation of GWAS findings to specific genes, proteins, or networks is frequently absent. It would be beneficial to link different datasets to develop a more complete picture of how the genetic variant results in an alteration to the phenotype. However, even at the level of gene expression microarrays, it is generally accepted that only data using a common platform can be integrated reliably in a quantitative manner (Chen et al., 2007). Integrating multiple datasets to quantitatively relate genomic, transcriptomic, and proteomic signatures that have used different sample preparation methods, technical approaches and procedures, analytical tools, and annotation methods, is extremely challenging.

There are several important considerations to bear in mind. One complication is that the effects of gene regulation, like the diseases themselves, are often tissue or even cell-selective, and involve tissue/cell-specific molecular networks (Lukk et al., 2010; Melé et al., 2015; Ni et al., 2016; Price et al., 2015). However, identifying the actual cells that are relevant, aside from identifying appropriate cell-specific data in the database, is not always clear-cut. As exemplified for coronary artery disease (CAD) by Price et al. (2015), it might be necessary to consider liver/hepatocytes involved in lipoprotein metabolism, cells of the immune system implicated in the development of CAD, cells of the blood, such as platelets and leukocytes, aside from the more obvious endothelial and smooth muscle cells of the artery wall itself.

While several thousand genes show tissue-preferential expression, only about 200 genes are expressed exclusively in a given tissue (Melé et al., 2015). The primary separation is between blood and solid tissues, and between solid tissues, the brain is the most distinct, within which the cerebellum is the most clearly differentiated (Melé et al., 2015). Interestingly solid tissue cell lines show similar gene expression patterns to each other and are distinct from their tissues of origin, but with close similarity to blood cell lines (Lukk et al., 2010), suggesting they are not good models of human tissues. Melé et al. (2015) identified 1993 genes that globally change expression with age, 753 with tissue-specific sex-biased expression, and 31 with tissue-specific ethnicity-based expression, predominantly in the skin. Failure to consider these factors when attempting to correlate gene expression to molecular networks that regulate the phenotype can obviously lead to erroneous conclusions.

1.5.4 Analytical Tools

These large datasets bring unique analytical and statistical challenges (Fan et al., 2014). The plethora of analytical tools for these large datasets is overwhelming, is constantly being

revised, updated, and changed, discussion of which could fill a book. What is disturbing is the extent to which the chosen analytical methods can impact the results and conclusions (Barrows et al., 2010; Clooney et al., 2016). For example, an analysis of the microbiome in human stool samples found that the chosen analytical method was responsible for more variance in gene expression than differences between the species of microbiota (Clooney et al., 2016). Updated analytical tools lead to different interpretations than those used when the data were first generated and deposited, but there is little incentive to go back and reanalyze and correct historical data. There is a concern that multiple analytical methods can be employed until the answer that is sought can be found, analogous to *p*-hacking to find a statistical test that shows a "significant effect" in other types of experiments (Simmons et al., 2011).

1.6 THE REPRODUCIBILITY PROBLEM

Research is defined as "the systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions" (https://en.oxforddictionaries.com/definition/research). A more precise definition applicable to biomedical research "is the broad area of science that involves the investigation of the biological process and the causes of disease through careful experimentation, observation, laboratory work, analysis, and testing" (CBRA, 2017) to which the authors would add "to identify the molecular causes of disease in order to identify drug targets that can be used to develop safe and efficacious therapeutics to benefit patients."

For a variety of reasons, it has been historically unusual for researchers to submit a study for peer review that is an attempt (or failed attempt) to repeat one already published in the literature and equally unusual for a journal to publish one (Buntin et al., 2011; Dirnagl and Lauritzeb, 2010; Fanelli, 2012; Ioannidis, 2005; Ioannidis and Trikalinos, 2005; Song et al., 2013; ter Riet et al., 2012; Tsilidis et al., 2013). The reproducibility of the original finding is therefore not commonly reported even when it undergoes evaluation. This is referred to as the "file drawer" effect and represents a publication bias where "journals are filled with the 5% of the studies that show Type 1 errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results" (Rosenthal, 1979). This bias runs the risk of publishing the 1 study out of 10 that showed a desired effect, and ignoring the other 9—both by a single investigator as well as other researchers who tried to reproduce the study (Begley, 2012).

The publication of negative outcomes occurs most frequently under a limited number of scenarios, for example, when:

- The new finding is potentially important: for example, a new disease mechanism or putative drug target, a critical methodology (Obokata et al., 2014a,b) or when a drug candidate enters Phase IIa clinical studies. In these instances, the stakes are high and external scrutiny becomes far greater, and where the inability to reproduce findings in preclinical research, often in animal models, leads to failed translation in the clinical setting—generating considerable concerns especially in the biopharmaceutical sector (Section 1.9).
- A systematic review of the type pioneered by the Cochrane Library, or a metaanalysis of multiple studies on a topic are compiled and analyzed, and can lead to conclusions that differ from many of the individual component studies and influence patient standard of care.

- The *Proteus Phenomenon.* It has been observed that the more extreme a research finding, the more likely it is to be statistically significant and published quickly (Ioannidis and Trikalinos, 2005). As discussed in Chapter 2, exploratory studies that are underpowered are more susceptible for producing outlying results that are not true and are difficult to reproduce. Consequently, further studies addressing the same topic might yield results in the same direction, but of a much lower magnitude, and hence less provocative and less attractive for publication. Since the initial publication represents results that are disproportionate to reality, it is equally possible that a follow up study could find the opposite results. Highly contradictory outcomes are more attractive for publication as they foster interest in determining the "right" answer, and potentially increase journal Impact Factors (Drucker, 2016). This bias in publication might, in large part, underlie the frequency with which dramatic findings in key areas of topical interest are frequently followed by equally dramatic findings in the opposite direction—an event termed the *Proteus phenomenon*, named after the Greek god able to assume multiple forms (Ioannidis and Trikalinos, 2005). Further studies then fall in between these two extremes, a regression to the mean identical to that described in clinical trials.

  While examples of the *Proteus phenomenon* have been published (Ioannidis, 2006; Ioannidis and Trikalinos, 2005; Pfeiffer et al., 2011), most of these relate to genetic markers in molecular medicine—a relatively easy target since genetic associations are low in the reproducibility hierarchy. What is much less clear is the extent to which this is a widespread phenomenon, its full impact on issues of reproducibility, and whether or not it is a result of publication bias, sloppy preliminary (exploratory) studies, or a combination of issues. Ioannidis and Trikalinos (2005) warned that the increased availability of large databases for interrogation by different groups simultaneously might result in increased contradictory publications occurring almost contemporaneously and increasing confusion, although whether that occurred in the ensuing decade has not been reported.

On a broad scale, the validity and reproducibility of data emerging from preclinical biomedical research has become a cause for concern (Begley and Ioannidis, 2015; Collins and Tabak, 2014; Dolgin, 2014; Jarvis and Williams, 2016; Prinz et al., 2011) irrespective of whether these data originate in academia or industry. Consequently, for the better part of the past decade, there have been numerous articles in the research literature and mainstream media on the shortcomings in the quality control processes in scientific publishing. These include peer review (Chapter 5.2) and scientific self-correction (Alberts et al., 2015; Kraus, 2014), both of which have been deemed inadequate in their current forms (Begley and Ellis, 2012; Estes, 2012; Flier, 2017; Ioannidis, 2005; Prinz et al., 2011). Separately, concerns regarding the relevance of modern day biomedical research to human disease (Horrabin, 2003) and a lack of efficiency in its execution (Chalmers et al., 2014; Macleod et al., 2014) have led to estimates that 50% of research, preclinical and otherwise, cannot be reproduced (Button et al., 2013; Goodman et al., 2016; Horton, 2015). Together these phenomena have given rise to concerns about the sustainability of the current biomedical research model in an era of relative economic austerity (Alberts et al., 2014; Balch et al., 2015).

Triaging the landmark Ioannidis, Prinz et al., and Begley and Ellis articles, The (Economist, 2013a) published a high profile, widely read, and widely cited report under the rubric, "Unreliable Research" that focused on the issue of reproducibility. This article also

highlighted additional concerns regarding peer review of research articles (Chapter 5.2–5.4), and the very low standards of scrutiny and output in the fringes of the scientific literature (Beall, 2012; Bohannon, 2013; Butler, 2013), which has undergone an exponential increase over the period 2010–14 representing 25%–30% of all published articles (Chapter 5.6.4) (Beall, 2016; Bohannon, 2015; Shen and Björk, 2015). The relevance of fringe research publication to the core rubric of science is small, but its overt visibility due to illogical, often Pollyannaish claims that are supported by minimal and/or low quality data is attractive to the mainstream media, and newsworthy, leading to its widespread dissemination. The worst types of fringe publications are those in so-called "predatory" journals, defined as "counterfeit journals to exploit the open-access model in which the author pays" (Beall, 2012). Predatory publishers generally lack transparency and any notion of peer review, and are considered as dishonest. Beall notes, "They aim to dupe researchers, especially those inexperienced in scholarly communication….[by setting up]… websites that closely resemble those of legitimate online publishers, and publish journals of questionable and downright low quality. Many purport to be headquartered in the United States, United Kingdom, Canada or Australia but really hail from Pakistan, India or Nigeria. Some predatory publishers spam researchers, soliciting manuscripts but fail to mention the required author fee. Later, after the paper is accepted and published, the authors are invoiced for fees, typically US $1,800. Because the scientists are often asked to sign over their copyright to the work as part of the submission process (against the spirit of open access) they feel unable to withdraw the paper and send it elsewhere." A recent development in the predatory publisher sphere is the disappearance of Beall's controversial blog site at the beginning of 2017 (Chawla, 2017). Beall noted that "he was worn out by publishers who threatened him and harassed his colleagues" (Gillis, 2017) and by "intense pressure" from his university about the list that led to him "fearing for my job." This led Beall (2017) to document the pressures brought to bear on his activities by irked predatory publishers and, in some instances, colleagues in the library sciences (Beall, 2017). This article also includes Beall's personal concerns regarding open access publishing and the current state of scientific publishing—a "once-proud scholarly publishing industry is in a state of rapid decline.. [with]... predatory publishers pos[ing] the biggest threat to science since the Inquisition."

Beall's list has been replaced by a "more transparent" subscription-based list, Cabell's Journal Blacklist (https://www.cabells.com/about-blacklist) that uses 65 criteria to determine whether a journal is "deceptive", this term replacing predatory as a descriptor (Gillis, 2017).

Major efforts are ongoing to find effective and logical remedies to the basic reproducibility problem, that are focused on good research practices involving improvements in experimental design, analysis, and disclosure and reporting (Begley and Ioannidis, 2015; Collins and Tabak, 2014; Curtis et al., 2015; Jarvis and Williams, 2016; Mullane et al., 2015; Wadman, 2013) together with improvements in *Classical Peer Review* (*CPR*; Chapter 5.3.2.1) that are intended to identify research that is credible, distinguishing it from that which is not.

1.6.1 Factors Contributing to the Reproducibility Problem

Among the factors proposed to contribute to the reproducibility problem are:

1. inadequate standards of training and mentoring of new generations of scientists (Flier, 2017);
2. an inability to repeat an experiment due to a scarcity of resources (Alberts et al., 2014; Balch et al., 2015; Goodstein, 1995) or adequate experimental detail (Vasilevsky et al., 2013);
3. institutional culture and pressure–or lack thereof (Chapter 6.6.2; NASEM, 2017);

4. distorted reward systems for researchers—often termed "perverse incentives"—that lead to hubris and fraud (Alberts et al., 2015; Begley and Ioannidis, 2015; Ioannidis, 2014; Smaldino and McElreath, 2016); and
5. detrimental research practices which, in many respects, parallel the ongoing ethical malaise in 21st century society (Bishop, 2013; Mullane and Williams, 2015; Redman, 2015).

1.6.2 Human Contributions to the Reproducibility Problem

Changes in cultural norms in biomedical research in the 21st century frequently reflect personal agendas that are prioritized to the absolute detriment of the greater good, eroding the commitment, inquisitiveness, and altruism that has been the cornerstone of the biomedical research culture that has sustained its success (Begley and Ioannidis, 2015; Ioannidis, 2014; Kraus, 2014). This has led to detrimental research practices that are enabled by the "perverse incentives" (Alberts et al., 2014; Begley and Ioannidis, 2015) provided by the structure of 21st academic biomedical research, the main currencies of which are citations and grants. These dictate the outcomes of grant applications and career advancement such that Richard Harris, the author of *Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions* has noted (Harris, 2017, p. 3) that "scientists often face a stark choice: they can do what's best for medical advancement by adhering to rigorous standards of science, or they can do what they perceive is necessary to maintain a career in the hypercompetitive environment of academic research."

Historically, scientific research has been enabled by philanthropists and by initiatives in federal research funding (Chapter 6; Triggle and Williams, 2015) after World War II. In the United States, this was part of the *Endless Frontier* (Bush, 1945) and in the United Kingdom, Butler's Education Act (Barber, 2014), both of which enabled access to higher education for a broader portion of the population. Thus, from the 1950s to the early 1970s, these initiatives provided funding to biomedical researchers with an expectation of infinite career growth, while also catalyzing the tools of molecular biology discovered by Watson and Crick (DNA), Milstein and Kohler (Monoclonal Antibodies), Mullis (PCR—polymerase chain reaction), and Boyer and Cohen (DNA cloning). These techniques aided in increasing the precision and, when accurate, the relevance of research outcomes, with the potential to revolutionize drug discovery. The flip side of the coin was that these sophisticated tools tended to become the focus of research activities with questions about human health becoming incidental or irrelevant (Horrabin, 2003). This prompted a reductionist, hypothesis-driven but poorly validated approach to biomedical research that is continually superseded by new tools with even greater precision and promise, but that, until recently, moved farther and farther away from the patients they were intended to help.

Two contrasting events in the 1980s dramatically changed the biomedical research landscape. First, the Bayh-Dole Act of 1980 (Boettiger and Bennett, 2006; Markel, 2013) gave universities ownership of their discoveries made using federal funds, with the obligation that the university should seek to license and commercialize such inventions. Second, the promises of the tools of molecular biology were beginning to be realized, with the first biologic products and monoclonal antibodies reaching the marketplace. This triggered interest from Venture Capitalists to invest in a fledgling biotechnology industry (Chapter 6.4.1.1). The two events together also provided an environment that allowed individual scientists to retain their research interests and, often, their academic positions, while having the opportunity to become entrepreneurs, starting their own companies, making themselves and their institutions wealthier (Kotecki, 2016; Vasagar, 2001). Inevitably, this dual role has created biases—perceived or real—in data generation, evaluation, analysis, and sharing, with inherent conflicts

of interest that compromise the integrity and value of information that is put into the public domain. There are also many interesting tales of rogue entrepreneurial scientists involved with insider stock trading and swindles that were supported by fraudulent, misinterpreted or nonexistent data. These have resulted in fines and, in some rare instances, prison sentences (Hodgson, 2016; Interlandi, 2006; Nutt, 2016; Prud'Homme, 2004).

Increased competition for funding and positions has, however resulted in aspects of biomedical research becoming self-serving, involving a substantial flow of taxpayer funding via federal grants to support researchers to conduct research, the main purpose of which appears to be to consolidate a position from which to apply for more grants, *ad infinitum, ad nauseam*. In some instances, the science becomes secondary to the pursuit of personal gain and career advancement whereby some researchers build a biotechnology company (Shimasaki, 2014) or aid in the commercialization of academic research which, by virtue of the stakeholders and the funding process, have been viewed as little more than Ponzi schemes (Mirowski, 2012).

While the onus to ensure integrity and maintain high scientific standards rests primarily with the individual researcher, nonetheless, institutes and the scientific community at large also have key roles to play. The NASEM (2017) report highlights the pressures many academic researchers face where an increasing number solely depend on grant funding for salaries (with all the attendant vagaries of federal support), and are required to maintain a publication record in high profile journals; and where the growing number of graduate and postdoctoral students are competing for an ever-diminishing supply of faculty positions. The report stresses the need for the scientist to "be free from pressures and influences that can bias research results" and compromise objectivity. Separately it may be argued that there is a mandate for the scientific community to recognize, acknowledge, and take responsibility for improving those aspects of the current biomedical research culture that reflect poorly on its societal role and that have the potential to further undermine its credibility with the public and result in cuts in funding (Alberts et al., 2014; Kornfeld and Titus, 2016; Nature Medicine Editorial, 2016). Although the shortcomings in reproducibility more often than not involve honest mistakes—faulty design, biased analysis, and the inherent vagaries of biomedical research (Flier, 2017; McClain, 2013)—this is frequently not obvious to the public. However, society clearly can understand deliberate fraud and there is a perception that it is the fraud that affects societal attitudes to science rather than shortcomings in design and analysis. Thus, as noted by Broad and Wade (1983) more than 30 years ago, "no matter how small the percentage of scientists who might be fakers of data, it require[s] only one case to surface every few months or so for the public credibility of science to be severely damaged." In many respects, outright fraud is merely the tip of an iceberg of the multiple reasons for failed data reproduction. The real concern is how much of the iceberg is due to lack of training, lack of authentication/validation of experimental materials, and lack of insight and knowledge that leads to poor experimental design and analysis, selective reporting, etc., and how much of it is due to more insidious issues, such as deliberate fraud, plagiarism, intrinsic incompetence, and hubris (Alberts et al., 2014; Begley and Ioannidis, 2015; Ioannidis, 2014; Nisbet and Markowitz, 2014; Redman, 2015; Tharyan, 2012).

An additional factor that confounds risk minimization is that many scientists cannot conceive that a colleague in their particular field of research would engage in any activities that are contrary to the pursuit of good science. This includes the routine use of substandard design, execution, and analysis of experiments, or indulging in plagiarism, or outright fraud. Instead, unacceptable behaviors are downplayed or even ignored, thus undermining the

bedrock of biomedical research (Stroebe et al., 2012). Accordingly, the biomedical research community must acknowledge the possibility of fraud, data manipulation, and the malicious intent of some authors. Indeed, deliberate fraud (Pulverer, 2016) is thought to be far more widespread than what the "research elites" that manage scientific publication and funding are currently prepared to admit (Kornfeld and Titus, 2016; Wilson, 2016).

Scientific progress is based on the notion that false premises are corrected by the emergence of robust and compelling new data that not only contradict the "state of the art" but aid in reconciling the past and the present by identifying honest misconceptions. This of course assumes that all misconceptions are the product of misapprehension, not flawed design, analysis, or interpretation. Given the discussion earlier, this can no longer be considered as a given in a society where ethics are undervalued or ignored (Bishop, 2013) and where the consequences for errant behavior are arbitrary, and often minimal (Kornfeld and Titus, 2016; Triggle and Williams, 2015).

1.6.2.1 Litigation in Basic Biomedical Research

Another change in biomedical research, again reflective of 21st century society, is the trend for litigation whereby researchers accused of fraud or other forms of scientific misconduct seek legal recourse. Unfortunately, this can dissuade institutions, funding bodies, and whistleblowers from engaging in prosecuting fraud and, indeed, even seeking to uncover fraud, and therefore to fail to take an appropriate course of action. Fear of retribution, being accused of sharing culpability for the fraud, and/or being sued can make it hard for rogue researchers to be identified publicly, resulting in questionable publications retaining, unchallenged, their currency and influence. Complaints from whistleblowers—usually unidentified individuals—are routinely dismissed as "a difference of opinion"—and can compromise careers, especially when the whistleblower is junior to the person who is the subject of their complaint (Goldberg, 2015; Nutt, 2016). Instead, the onus is often on the whistleblower. Some whistleblowers have had a major impact in improving institutional research standards, but many are harassed by the administration at their institution or have their concerns repeatedly dismissed despite having spent considerable time and personal resources in order to gather data to prove their case, while others are simply ignored (Gross, 2016; Yong et al., 2014).

Even when papers are retracted, the authors often continue to "win" grants and publish new research. Often, all that is required to retain a modicum of credibility is a claim by the principal investigator or laboratory head that the false data were included "by mistake." This is normally accepted, especially by publishers who fear litigation if they were to "blacklist" an author. Like big business, accused individuals and their institutions will often pay a fine or offer a short statement of retraction to avoid prosecution while not admitting "any liability or wrongdoing." Instead of being the subject of professional or criminal investigation, they are given a slap on the wrist, for example, being barred from applying for grants for 3 years (Kornfeld and Titus, 2016; Nutt, 2016). However, the arbitrary and capricious nature of penalizing fraudulent activities is evidenced by the laboratory of fired NIH neurologist, Allen Braun where all the researchers have been barred from publishing their data, including graduate students and postdocs who need the publications to embark on their careers, demonstrating that collateral damage can extend to associates not accused of any misdeeds (Couzin-Frankel, 2017). Contrast that outcome with the notorious case of John Darsee, a cardiologist at Emory and Harvard who fabricated much of the data contained in over 100 publications with 47 coauthors. While Darsee was fired and the coauthors could lay claim to being

hoodwinked by his activities, nevertheless a review of 18 of his retracted papers (Stewart and Feder, 1987) revealed a catalog of serious flaws (besides whether the data were valid), including reuse of historical control data in multiple articles without even acknowledging the fact; publication of the same data in different journals under different titles (with four coauthors in common and where Darsee was neither the first nor the last author); multiple cases of inconsistent values between the text and figures and standard deviations that did not match the datasets; issues so egregious that the coauthors must have been aware, and in some cases actually defended Darsee (Stewart and Feder, 1987), and where there is no evidence that the coauthors were ever brought to task over the incident.

The lack of significant penalties for scientific fraud also extends into the corporate world of biopharmaceuticals where the management of a number of companies are widely known to have selectively reported clinical data that presented the company in a favorable light, while ignoring negative data to enhance stock value. Much of this information is murky in origin, open to interpretation, and certainly gist for corporate lawsuits against those making such assertions. One such entity is the "stealth" clinical diagnostic company, Theranos, that has been the subject of numerous articles in the mainstream media describing the lack of transparency in its technology. Historically, Theranos submitted very little, if any, of its purported breakthrough technology for peer review (Bilton, 2016a; Carreyrou, 2015; Ioannidis, 2015) with much of their research apparently being "doctored," having failed quality-control checks (Bilton, 2016b; Carreyrou, 2016). Indeed, Theranos is reported to have actually set up a secret company to buy commercial clinical devices in order to run "fake 'demonstrations tests' for prospective investors and business partners" while pretending to showcase its own technology (Weaver, 2017). As a result, the US Centers for Medicare and Medicaid Services, in addition to revoking the license of the company to operate a clinical testing laboratory in California because of unsafe practices that posed an "immediate jeopardy to patient health and safety," banned the company's founder from the blood-testing business. But only for 2 years: a questionably lenient penalty (Flam, 2016; Scott, 2016) that may be superseded by numerous investor lawsuits (Kossoff, 2017).

Because of the risk of litigation, many questionable research findings have sat in the publication cloud, unmolested, despite unspoken (and sometimes open) criticism, and skepticism regarding their provenance. Some of these articles will have been unfairly maligned while others that are truly fraudulent may have maintained a baleful influence. This miasma has ameliorated only somewhat with the advent of the website, *RetractionWatch* (http://retractionwatch.com).

1.6.3 The Impact of the Internet on Publishing and Disseminating Information

With the advent and spread of the Internet, the process of scientific publication has irrevocably changed, becoming faster, more transparent, and interactive. Identifying and recruiting reviewers has similarly become a rapid, semiautomated process. This has allowed the *CPR* process to accept papers for publication within days of submission and their publication online within days of acceptance. Software development has enabled the emergence of *postpublication peer review* (PPPR; Chapter 5.3.2.3) and *postpublication commentary* (*PPC*; Chapter 5.3.2.4). These, while well intended, are insufficiently well curated such that they enable aspects of "vigilante science" that serve no useful purpose (Blatt, 2015). PPPR and *PPC* are thus highly

vulnerable to "'trolling'—the posting of disruptive or malicious comments" (Stoye, 2015). PPPR and *PPC* have had limited impact despite extensive proselytization by their advocates that they are a quantal improvement compared to the traditional peer review process.

Nevertheless, the Internet publishing model, together with the proliferation of research papers submitted and published, contrasts markedly with the era of paper publishing with its relatively slow peer review process and perhaps, a greater and more thoughful degree of due diligence. Internet publishing has also contributed to more papers having a shorter "shelf-life" with many rapidly disappearing "without trace" in the virtual sea of the 1.4–2 million papers that are indexed in PubMed *each year*.

Similarly, the large number of journals with low levels of peer review standards commensurate with their low, or nonexistent Impact Factor offers multiple alternative forums for dissemination such that a submitted manuscript will eventually find a journal that will accept it, a trend in publication described by Peres-Neto (2016) as "where rather than if." Separately there are the predatory journals, many of which have published nothing, and which lurk in cyberspace, awaiting the naive researcher only too happy to be parted from their grant money, especially with a publication at the end of the interaction. These journals are of dubious merit for authors who use them who are, as noted, naïve or work in an environment (possibly one involving the scientific culture of a complete country; Economist 2013b) that lacks the appropriate ethical standards and judgment necessary to recognize their lack of value. Although many researchers working in institutions in countries with a long and distinguished record of research publication may regard the predatory platform as irrelevant to their activities, the fact that these journals exist and have proliferated attests to the existence of a market, and one that is growing exponentially. An eightfold growth in the number of papers published in the predatory literature over the period 2010–14 (Shen and Björk, 2015) has increased the contribution of this sector to an estimated 20% of the total published scientific literature in 2015 (Ware and Mabe, 2015).

1.6.3.1 High and Low Profile Journals

Another component of the reproducibility problem is the large number of papers published in the biomedical research literature that are rarely cited; instead they are "sitting in a wasteland of silence, attracting no attention whatsoever" (Davis, 2011). As a result, the findings reported in these papers are rarely subjected to independent verification. This may be because they reflect a research area that is a scientific backwater or because they are published in a relatively obscure journal (Schmidt, 2014). On the other hand, it may be because they are too far ahead of their time or because they show a hypothesis to be false and thus terminate an interesting and easily fundable avenue of research. These explanations do not imply that there is anything wrong with the work. However, a lack of independent verification may be because the research community assumes the data are false for unknown reasons and choses to ignore the publication. This assumption may or may not be correct and may be the result of the journal itself having a low Impact Factor (an example of the publishing sector vicious circle), resulting in a readership bias. In contrast, papers that are published in high profile, high Impact Factor journals like *Cell*, *Nature*, or *Science* attract interest and have inherent "credibility" owing to the "marque" of the journal, so readers are more inclined to assume that the findings are likely to be correct. These inspire attempts to reproduce the studies which, when they cannot be reproduced, immediately have a high profile which is why high Impact Factor journals have high retraction rates. Fang and Casadevall (2011) have reported a positive

correlation between the journal Impact Factor, the visibility of a journal, and the number of retractions. This may seem counterintuitive, since one would expect the high Impact Factor journals to be publishing papers that are more likely to be correct although their peer review mechanisms appear no better than those in lower impact journals. Instead, high profile journals are more likely to publish cutting edge and impactful research—research that attracts more attention—and, consequently more attempts to reproduce the findings than research published elsewhere (Bishop, 2012; Fang and Casadevall, 2011). This is a perception facilitated by the public relations efforts of the high impact journals to ensure that their output is avidly awaited and disseminated by the mainstream media, sometimes even before the paper documenting the research is officially published.

The corollary of this is that there is a greater chance of "getting away" with publishing false or fabricated findings, if an author opts for a lower impact journal (and also focuses on more mundane research issues) as few will read the work, fewer will regard it as important, and yet fewer still will attempt to replicate the findings.

1.7 THE LITERATURE ON REPRODUCIBILITY IN BIOMEDICAL RESEARCH

1.7.1 An Age-Old Concern

Concerns related to reproducibility in biomedical research are far from new, despite the apparent surprise (Economist, 2013a) that greeted the papers from Prinz et al. (2011) and Begley and Ellis (2012). For example, the reproducibility of published findings formed part of the feasibility assessment for initiating a new project, and was *de rigueur* throughout the pharmaceutical industry for many decades before the Prinz et al. (2011) paper was published.

When original research findings could not be reproduced, it was a standard operating procedure to contact the original authors to resolve any potential oversights or methodological disconnects. This discussion often led to a satisfactory resolution of the problem by changing reagents or protocols or clarifying procedures. When the cause of failed reproducibility remained elusive or unresolvable, the outcome of the study which failed to reproduce the original findings was very rarely published but its existence became well known in both academia and industry by way of informal research networks (chatter and gossip), long before the advent of the Internet. Thus, when the research sector was small, controversy was rarely a secret, and the chance of unverifiable findings being accepted as fact was lower than it is today—in large part because the number of scientific publications was only a fraction of today's output.

1.7.2 Concerns in the 21st Century

Prior to the Prinz et al. paper, concerns with data reproducibility had been expressed by Ioannidis (2005) in his seminal article "Why most published research findings are false" on experimental reproducibility that addressed the probability that the majority of findings in research were false positives and has been accessed well over a million times. This article was discussed in two articles in the mainstream media, one entitled "Lies, Damned Lies, and Medical Science" (Freedman, 2010) in the *Atlantic Monthly* and the other, "The Truth Wears Off" (Lehrer, 2010) in the *New Yorker*.

The Freedman piece focused on the metaresearch being conducted by Ioannidis and his colleagues discussing many types of bias (including data selection and data analysis, exciting rather than plausible theories, etc.) that distorted research outcomes and led to the publications that were "pervasively flawed…misleading, exaggerated, and often flat out wrong" and that frequently contradicted previous reports in the peer-reviewed scientific literature.

The Lehrer report also cited the contributions of Ioannidis et al. to the reproducibility debate extending the discussion to the phenomenon known as the "decline effect" which has been discussed in detail (Section 1.4.3). This term had been originally used in the psychological sciences to describe how early reports of evidence for significant extrasensory perception (ESP) in experiments on the phenomenon of "psychic" powers and has since been used as a general term to describe a situation where initial experimental results are highly impressive, for example, significant at $p < 0.05$, etc., but with time become less significant as investigators attempt to reproduce the original findings.

Thus, the novelty in the Prinz et al. (2011) paper was its specific focus on *numerous instances* of a failure to reproduce findings from activities conducted in a single research center that was dedicated solely to target validation as well as its high profile and graphically colorful venue, *Nature Reviews Drug Discovery* (Jarvis and Williams, 2016).

An inevitable question, given the widespread response—and, in some instances, outrage—to the Prinz et al., and Begley and Ellis papers, is whether scientific misconduct is on the increase or simply reflects the fact that the means to detect its occurrence has become more efficient via the Internet. Data suggest that the approximate 1% incidence rate for scientific misconduct in published articles has remained unchanged for more than a decade (Blatt, 2015).

1.7.3 Reproducibility in the 21st Century: Origins, Scope and Momentum

1.7.3.1 Why Most Published Research Findings are False (Ioannidis, 2005)

Most published findings in biomedical research are positive, reflecting a publication bias in favor of authors submitting, and reviewers accepting, positive versus negative findings (Dirnagl and Lauritzeb, 2010; Fanelli, 2012; Ioannidis, 2005; Ioannidis and Trikalinos, 2005; ter Riet et al., 2012; Song et al., 2013).

In this article, Ioannidis identified a number of factors that contribute to the false positive rate including: the statistical power of a study; the level of statistical significance; bias as represented by the selective or distorted reporting of data (described by Babbage in the 19th century as "trimming"; Gross, 2016); prejudice and financial interests; and effect size where the tested relationships were greater in number and had not all been preselected as endpoints. As a result, this leads to a greater flexibility in "designs, definitions, outcomes, and analytical modes" which can be interpreted in terms of experiments becoming infinitely variable—changing in their substance and endpoints from the conceptualization of the study to its interpretation for a variety of reasons including a lack of rigor and/or knowhow and/or bias/honesty.

Couched in language and equations that, despite their intended transparency, were challenging to the statistically naïve, especially those unfamiliar with the concepts of Bayesian statistics, Ioannidis' paper focused on the aspects of statistical issues, particularly the relevance of the "ill-founded 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." With the further revelation that "Research is not most appropriately represented and summarized by *p*-values" many biologists who were taught the infallibility of the *p* value understandably became nervous and confused by this, and subsequent papers from the prolific Ioannidis. Whether data are sufficiently "good" to justify that the *p* value is a different issue and is discussed further in Chapter 2 along with the *null hypothesis.*

1.7.3.2 Believe it or not: How Much can we Rely on Published Data on Potential Drug Targets? (Prinz et al., 2011)

Prinz et al. (2011) was the first of the two papers from the pharmaceutical industry that served as catalysts to highlight concerns regarding reproducibility. It originated from the German pharmaceutical company, Bayer HealthCare and reported that 65% (43/69) of findings published in the literature identifying targets in the areas of oncology, women's health, and cardiovascular disease could not be reproduced when subjected to internal validation efforts. This evaluation was extended to an internal Bayer survey of 23 colleagues across different scientific disciplines, the results of which indicated that only 20%–25% of the published data used by the Bayer project teams to validate projects could be reproduced in house. While citing an "unspoken rule" in venture capital circles (https://lifescivc.com/2011/03/academic-bias-biotech-failures/) that "at least 50% of published studies, even those in top-tier academic journals, can't be repeated with the same conclusions by an industrial laboratory," Prinz et al., did not conclude that the original findings were incorrect or ambiguous due to deficiencies in the reported experimental parameters (e.g., cell line authenticity, assay formats). Rather they emphasized inappropriate experimental design and analysis (deficiencies in null hypothesis testing, inherently low prestudy probabilities of observed results actually being true), associated with small sample sizes that reflected the pervasive "publish or perish" scientific environment, as the primary contributing factors. The original papers also appear to have overlooked the almost ubiquitous lack of blinding and randomization in the preclinical research area (Hirst et al., 2014; Chapter 2). Of additional note given the high profile of the Prinz et al., report, the data on which it was based did not identify the specific targets involved thus making the findings uncorroboratable while the primary publications being assessed for reproducibility were also not identified (Jarvis and Williams, 2016), making it impossible to verify the findings.

1.7.3.3 Drug Development: Raise Standards for Preclinical Cancer Research (Begley and Ellis, 2012)

A second industry report on reproducibility originated from the Californian biopharmaceutical company, Amgen and the MD Anderson Cancer Center in Houston, Texas, a major US cancer research institute. This paper focused exclusively on "landmark studies" in preclinical cancer research. Of these, the findings in only 6 of these studies in a cohort of 53 could be reproduced in house. While this was concerning in itself, the responses received in following up with the original authors generated additional, if not greater, concern. Thus, Begley and Ellis stated that "To address these concerns, when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors' direction, occasionally even in the laboratory of the original investigator. These investigators were all competent, well-meaning scientists who truly wanted to make advances in cancer research. In studies for which findings

could be reproduced, authors had paid close attention to controls, reagents, investigator bias, and describing the complete data set. For results that could not be reproduced, however, data were not routinely analyzed by investigators blinded to the experimental versus control groups. Investigators frequently presented the results of one experiment, such as a single Western-blot analysis. They sometimes said they presented specific experiments that supported their underlying hypothesis, but that were not reflective of the entire data set. There are no guidelines that require all data sets to be reported in a paper; often, original data are removed during the peer review and publication process." In a subsequent meeting "with the lead scientist of one of the problematic studies" that was summarized by Begley; "We went through the paper line by line, figure by figure," ….. "I explained that we re-did their experiment 50 times and never got their result. He said they'd done it six times and got this result once, but put it in the paper because it made the best story. It's very disillusioning" (Begley, 2012). Begley provided further background detail on the Amgen reproducibility efforts to Harris (Begley, 2017b) noting that "On about twenty occasions we actually sent [company] scientists to the host laboratory and watched them perform experiments themselves." These were blinded studies that "most of the time …failed."

The general conclusions in the Begley and Ellis paper were twofold: (1) that reproducible research tended to reflect a high level of attention to controls, reagent quality, and descriptions of complete datasets and; (2) research that could not be reproduced often lacked detailed descriptions of experimental methodologies and involved data sets that were representative in nature precluding testing of reproducibility due to vagueness.

As in the Prinz et al. (2011) report, the data in the Begley and Ellis paper were again largely unverifiable with none of the original papers being identified due to the execution of confidentiality agreements with some of the authors of the original reports (Nature, 2012), a situation that had not been viewed favorably (Gorski, 2012; Jarvis and Williams, 2016). This concern however can be viewed as a little disingenuous. Unless one can argue that Prinz, and Begley and Ellis fabricated their findings, which is unlikely, it is a little harsh to criticize them. After all, any criticism of this kind would argue that any study in which the identity of the subject matter or materials is anonymized must be regarded as illegitimate. Since the anonymity of human subjects is *de rigueur* in all ethically designed and reported clinical trials, then the acceptance of such criticism would argue for the abandonment of the current model for clinical research—which some have advocated (Grove, 2011)—a viewpoint that has garnered little in the way of support (Gorski, 2011; Lowe, 2011).

1.7.3.4 Reproducibility: An Academic Viewpoint

Prompted by the Prinz et al., and Begley and Ellis papers, Mobley et al. (2013) conducted an anonymous survey of academic researchers at the MD Anderson Cancer Center to determine the frequency and potential causes of nonreproducible results. From this survey, they reported that the problem of data reproducibility was well known in academia with some 50% of respondents having encountered at least one incident of irreproducibility in their career up to the date of the survey. Many were unable to identify its cause. When contacting the original authors to resolve their findings, responding researchers were met with "a less than 'collegial' interaction" from the original authors of whom "almost half responded negatively or indifferently." While individual responses to the survey were provided, the data in this paper were somewhat unrepresentative with only a 16% (434/2692) response rate (Section 1.12.1).

1.7.3.5 Reproducibility in the Psychological Sciences

An additional milestone in the evolving literature was the Reproducibility Initiative (RI) conducted by the *Open Science Collaboration* (OSC, 2015), a large-scale collaborative effort over the period November 2011–December 2014, to estimate the reproducibility of psychological science findings. This study, also known as *Reproducibility Project: Psychology* or *OSC2015* set the goal of replicating (e.g., reproducing) the findings from 100 published papers in the discipline of psychology that were published from 2008 onward in three leading journals in the field, *Psychological Science*, *Journal of Personality and Social Psychology*, and the *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Reproducibility was assessed on the basis of significance (reported p values), effect sizes, subjective assessments of expert teams, and metaanalyses. While the reader is directed to the original report for full details (OSC, 2015) and discussion in Chapter 5.12–5.14, the salient outcomes of the initiative were that of the 97% of the original studies that reported significant results ($p < 0.05$), only 36% of the repeated studies obtained significance. Thus 65% of the findings could not be reproduced (according to the impartial arbiter of statistical significance) even though some of the original authors were involved in the attempted reproduction effort. This appeared to indicate that reproducibility in the psychological science is low (Gibson, 2012; Gilbert et al., 2016; Stroebe et al., 2012) and led to the conclusion that statistically significant outcomes have an extraordinarily poor level of credibility given their poor level of reproducibility. There is no agreed reason for this, with flaws in design, analysis and basic scientific knowledge, and know-how all identified as culprits. This is discussed further in Chapter 5.12.

1.7.3.6 Researcher Awareness of Fraud

Four years before the Mobley report, Fanelli (2009) had conducted a metaanalysis of 21 surveys that had asked scientists about their experiences related to misconduct in science. The results—which Fanelli characterized as "conservative" given the sensitivity of the questions being asked—led to approximately 2% of scientists to admit having "fabricated, falsified, or modified data…at least once" with another 34% admitting to "questionable research practices." Of additional relevance was that 3 years prior to the Fanelli study, in 2006, the NIH Office of Research Integrity (ORI) had funded a Gallup survey looking into research integrity. The final version of the survey results was not released until 2008 and "estimated that 1.5% of all research conducted each year would be fraudulent," (Wells, 2008) consistent with Fanelli's later findings. Given that these findings were based on admission of guilt, this may be the tip of yet another iceberg.

Ten years later in a subsequent paper, *How to Make More Published Research True*, Ioannidis (2014) proposed that false positives in the biomedical literature would diminish if there was an increase in collaborative research, and an improved culture of reproducibility and replication with improved statistical methodology and user literacy with more stringent statistical thresholds, and improved study design standards for peer review, reporting, and dissemination of research. Added to this was the elimination of "nonmeritocratic practices" (Ioannidis, 2014) also known as "perverse incentives," improvements in the training of the scientific workforce and modifications in the reward system for science, with the latter being increasingly identified as dysfunctional and in need of restructuring (Alberts et al., 2014; Begley and Ioannidis, 2015; Ioannidis and Khoury, 2014; Kraus, 2014).

In addressing necessary changes in the historically nepotistic grant-funded ecosystem (Ioannidis, 2014), the adoption of which may take many years, especially when citation-based incentives uniformly select for bad science (Smaldino and McElreath, 2016), the absence of clear consequences and incentives to change entrenched behaviors (Fishburn, 2014; Nardone, 2008) represents a major shortcoming that no amount of wishful thinking (or additional grant funding) will overcome. In yet another paper, Ioannidis (2016a) distinguished clinical from preclinical or discovery research, terming the latter as "blue-sky" and "speculative" that unlike clinical research lacks prespecified deliverables but is nonetheless valuable. This is an important distinction as both the psychological and clinical sciences are often viewed as benchmarks against which preclinical research and its reproducibility should be marked. This has led to the viewpoint that preclinical experimentation should seek to emulate the gold standard of randomized clinical trials (RCTs; Henderson et al., 2013; Muhlhausler et al., 2013), another issue that is discussed in Chapter 2.

It is worthwhile noting that preclinical research is not always "blue sky" or exploratory, a quality that is certainly absent from the preclinical research dossier collated to support an IND (McGonigle and Williams, 2014), but should include properly powered, confirmatory studies. Finally, a point worth repeating *ad infinitum* and which is discussed in Chapter 2 is that the design of the experiment is a key to a good outcome. Deciding how to analyze an experiment *after* it has been completed is bad science (Chapter 3). Harking, hypothesizing after the results are known (Kerr, 1998; Motulsky, 2014), where multiple hypotheses are considered after review of the data and one is selected as though it were the basis for the study, is another source of irreproducibility since it not only introduces bias, but fails to account statistically for the multiple comparisons.

1.8 IS THERE A REPRODUCIBILITY CRISIS?

Approximately 4 years after the initial flurry of papers on reproducibility, a *Nature* survey involving 1,576 researchers found that more than 70% of respondents had tried and failed to reproduce an experiment (Baker, 2016b). And of this albeit self-selected and small cohort, 52% noted there was a "significant" crisis with another 38% noting a "slight" crisis, a view that is perhaps more in line with Gorski's, that "Reproducibility is… a chronic problem…. not a crisis" (Gorski, 2016). In responding to a question as to the cause of problems with reproducibility, more than 60% cited the pressure to publish and selective reporting which one respondent ascribed to the need to compete for grants coupled with an increased bureaucratic burden that "takes way from time spent doing and designing research." This is interesting given 21st century expectations for biomedical researchers, summarized by Faulkes (2016): "When I was in grad school, I had to write a paper and publish it. Now, people are suggesting that I also pre-register my experiments; curate and upload all my raw data (which may be in non-standard or proprietary formats); deposit pre-prints; publish the actual paper in a peer-reviewed journal (because that's not going away); promote it through social media; upload it into sites like Academia.edu or ResearchGate; update my publication information in databases like ORCiD, ImpactStory, and institutional measures; and watch for comments on post-publication peer review sites like PubPeer and engage with them as necessary."

Among the approaches suggested to improve reproducibility were: "more robust experimental design; better statistics; and better mentorship." Journal checklists (Chapter 5.11) received a 69% endorsement while 80% of respondents thought "funders and publishers should do more to improve reproducibility." This indicates that researchers are asking for help and lacking in the initiative and necessary know how to ensure that their own research is adequate—again reflecting disconnects in their training and mentoring.

1.9 TROUBLE AT THE LABORATORY?

An additional source of accessible, expert, and at times, provocative commentaries on topics related to reproducibility that include scientific publishing, peer review, and the integrity of the clinical trial process are certain publications in the mainstream media. These include the *Guardian* newspaper in the United Kingdom, and the *Economist, New York Times, Wall Street Journal*, *Forbes, Atlantic Magazine, Wired, Slate,* and *Vox*. Setting the tenor of these contributions was Colquhoun's frequently cited polemic in the *Guardian*,"Publish-or-perish: Peer review and the corruption of science" (Colquhoun, 2011) along with commentary by Goldacre in his book, *Bad Pharma* (Goldacre, 2012) and in his column/blog, *Bad Science* (http://www.badscience.net) that focuses on shortcomings in clinical trial execution in the pharmaceutical industry (with an often devastating and journalistically entertaining bottom line). Concomitant with this coverage has been an explosion of articles "researching research" (Ioannidis, 2016b) also termed metaresearch or metascience (Munafò et al., 2017) facilitated by both alternative (*Public Library of Science (PLoS)*, *PeerJ*, e*Life*) and traditional (*Nature, Science, Cell, PNAS, Lancet, etc.*) publishing venues as well as raw data access in formal databases. These have further highlighted reproducibility problems, and have proposed remedies—including the various *Reproducibility Initiatives* (Morrison, 2014; OSC, 2015; Van Noorden, 2014) and checklists for journals and grant submissions (Jarvis and Williams, 2016; Nature Editorial, 2016).

Additionally, there has been considerable discussion of the role of peer review, contested, long standing, still unresolved, and perhaps irresolvable (Colquhoun, 2011; Dzeng, 2014; Horrabin, 2001; Horton, 2000; Rennie, 1986; Rennie et al., 2003; Schekman, 2013; Smith, 2006; Souder 2011; van Dalen and Henkens, 2012), a topic that is covered further in Chapter 5.2.

1.10 RETRACTIONS

When a previously published effect cannot be repeated resulting in a *negative finding,* historically it has been very difficult—if not impossible—for a researcher to find a peer reviewed forum where this work will be accepted for publication. This is a consequence of the widespread bias in the peer review process toward positive findings (Buntin et al., 2011; Dirnagl and Lauritzeb, 2010; Fanelli, 2012; Ioannidis, 2005; Matosin et al., 2014; Pusztai et al., 2013; Song et al., 2013; ter Riet et al., 2012; Tsilidis et al., 2013; Young et al., 2008). Publication of contrary findings, a correction, or a retraction in the original journal in which the flawed finding was published provides a logical opportunity to be "complete and honest, and clearly articulate(s) what things…wrong" setting the record straight, making the investigator whole and avoiding the issue of citation penalties (Lu et al., 2013). When the originating author is

not the source of the subsequent failure to reproduce, the publication of the failed attempt to reproduce the finding can provide an opportunity for open discussion between interested parties and to avoid other researchers investing their time and resources trying to reproduce erroneous findings (Gewin, 2014). However, negative results have rarely been considered interesting or worthy of publication and certainly do not add to the "grant worthiness" of a biomedical researcher, except when it is related to a groundbreaking discovery and the *Proteus phenomenon* (Section 1.5).

The number of retractions—the public withdrawal of a published research paper—has increased markedly over the past decade (Boston, 2015; Steen, 2011a,b; Steen et al., 2013; Van Noorden, 2011), although it is still relatively small with estimates that vary from 1 in 10,000 (van der Vet and Nijveen, 2016), 1.4 per 100,000 (Marcus and Oransky, 2014) and "a tiny fraction of all published papers, perhaps several hundred of the one million published yearly" (Flier, 2017). While the majority of retractions were thought to be attributable to honest mistakes (Steen, 2011a), some two-thirds involve active misconduct that includes plagiarism, image manipulation, and other types of fraud leaving only 25%–33% being characterized as "honest mistakes" (Fanelli, 2009; Fang et al., 2012; Gewin, 2014).

Retraction, self or imposed, for example, a journal expression of concern or an outright rejection imposed by an Editor, is viewed as "the worst outcome of publication" (Souder, 2011). Thus, authors are not motivated to retract research papers because of the potential impact on their tenure, funding prospects, or prestige (Marcus and Oransky, 2014). In such instances, other scientists can express concerns via open blog sites like *PubPeer*. The lack of peer criticism on these sites however, leads to what has been termed "vigilante science" (Blatt, 2015) and dismissed accordingly. Concerned researchers and the editors of scientific journals can contact the authors and/or their institution when issues occur regarding papers with data that is suspected to be fraudulent, plagiarized, or cannot be reproduced, three very different degrees of concern. Taking these approaches, resolution can be achieved by the original journal publishing an expression of concern regarding a study or, when they have satisfactorily investigated a particular case, unilaterally retracting the paper.

When the authors and/or their institution are intransigent (Marcus and Oransky, 2014), negotiations can often become protracted and may involve legal complications (Lowe, 2011; *RetractionWatch*, 2016). This has consequences, especially for the researchers involved since it may result in loss of citations, not only for the retracted work but also the prior work of the "wronged" author (Lu et al., 2013) in the case of refutation or plagiarism. This is an occurrence that has been described as being "consistent with the Bayesian intuition that the market inferred their work was mediocre all along" (Azoulay et al., 2015).

1.10.1 *RetractionWatch*

A major resource in documenting irreproducible research, much of it leading to retractions of suspect papers that neither the authors nor the journals were motivated to do, is the website *RetractionWatch* launched by The Center for Scientific Integrity in 2010 (Carey, 2015; Gewin, 2014; Marcus and Oransky, 2011, 2014; Steen, 2011a,b).

RetractionWatch reports almost on a daily basis, new instances of suspected plagiarism, image manipulation, and fraud and the ongoing efforts to resolve these, all of which represents compulsive reading that is avidly covered in the mainstream media (Boston, 2015;

Carey, 2015). *RetractionWatch* has also collated a list of the more egregious offenders, the *Retraction Watch Leaderboard* (http://retractionwatch.com/the-retraction-watch-leaderboard/), where scientists receive scores for their retractions, and also publishes a yearly *Top Ten* list of offenders (Marcus et al., 2015). At the time of writing (Spring of 2017), two anesthesiologists, Fujii and Boldt, occupy the top positions on the Leaderboard with a combined 277 retractions. The retraction scores for these individuals were: Yoshitaka Fujii from Japan (Carlisle, 2012; Marcus and Oransky, 2015) leading with 183 retractions and Joachim Boldt from Germany (Wise, 2013) with 94. They were followed by Peter Chen, a Taiwanese physicist with 60, Diederik Stapel, a social psychologist from the Netherlands with 58 retractions (Stroebe et al., 2012; Verfaellie and McGwin, 2011), and Shigeaki Kato, a Japanese endocrinologist with 38 retractions. Another anesthesiologist, Scott S. Reuben (Bornemann-Cimenti et al., 2016; Borrell, 2009) from the United States was listed at position 9 with 25 retractions and was one of the unusual instances where a jail term was the outcome for his misdeeds (Kornfeld and Titus, 2016). Conclusions based on the data contained in the combination of retracted papers have had a major impact on the practice of anesthesia and also the postoperative treatment of surgically-related pain. It has been estimated that Rueben's fraudulent papers affected the treatment of millions of patients, and facilitated the sales of prescription COX-2 NSAIDs to the tune of billions of dollars (Borrell, 2009).

In a derisive letter to the Editor, Kranke et al. (2000) had noted in the title how "incredibly nice!" the data in Fujii's papers were. A belated *posthoc* analysis of a set of these papers indicated that the odds of some of them being actually experimentally derived were in the order of 10^{-33} which was viewed as "a hideously small number" (Carlisle, 2012). Despite the Kranke et al., letter, the 48 articles cited in the letter were not retracted and Fujii was able to publish another 11 papers in the same journal in the ensuing 10 years before a new editor took appropriate action (Stroebe et al., 2012).

Salient questions to explore are why the field of anesthesia is so susceptible to egregious fraud and numb to its consequences and whether the incidence of fraud in the area has increased in recent years. In considering the latter question, the editor of one of the journals involved, Shafer (2011) argued that the serial frauds perpetrated by Boldt and Reuben were far from new and that their personal misconduct dated back for over a decade, but failed to address this question. What was new however was the discovery which was facilitated by access to multiple sources of information available via the Internet. Conversely, others have wondered whether the field of anesthesiology is more vulnerable to research fraud because the individuals who self-select for this medical specialty may be more solitary by nature and receive less peer oversight than other medical subdisciplines.

As discussed further in Chapter 5.2, the peer review process has not always kept pace with changes in the scientific culture with an antiquated, entrenched code of conduct (Ioannidis, 2014) more appropriate to a bygone era than the cultural and ethical mores of the 21st century (Redman, 2015). The latter have been described in terms of a "Cheating Culture" (Callahan, 2004) that reflects "Norms [that] may arise within an organization that give implicit permission for unethical misconduct. A cheating culture exists when enough people are breaking the rules that there is a perception that 'everybody' is corrupt and there is no clear imperative for ethical behavior. In extreme instances, there may be the belief that one cannot be competitive by following formal rules and that cheating is the key to success" (Stone, 2005), analogous to the arguments made in sports by athletes who take performance

enhancing drugs. This would suggest that the necessary trust in others which has been assumed to be the core of scientific behavior and peer review (Kraus, 2014) has disappeared with the "trust me model that is no longer considered appropriate in corporate life nor in government" being extinct (Begley and Ioannidis, 2015).

Another variable deemed to be a major cause of fraud is a lack of basic rigor, competence, and expertize in the peer review process that is reflected in the finding that many of the retracted anesthesia papers lacked basic documentation including the mandatory Institutional Review Board (IRB) approval necessary to conduct studies in humans thus removing them from the administrative and peer physician radar.

1.10.2 Continued Citation of Retracted Publications

Retraction of inaccurate or fraudulent publications does not resolve the perpetuation of the inaccuracies that continue to be used and cited in secondary studies by other researchers. Of 82 retracted articles published between 1973 and 1987, there were 733 citations occurring more than 6 months after their retraction, a fall of only 18% compared to the preretraction period, with only 2.9% citations mentioning the retraction (Pfeifer and Snodgrass, 1990). Similarly examining 235 retracted articles from the period 1966 to 1996 that received 2,034 postretraction citations, only approximately 7% acknowledged the retraction, while all other citations used them with positive connotations (Budd et al., 1999). Moreover, most citations occurred in the Introduction or Discussion sections of the secondary publications, inferring that the retracted studies impacted the hypothesis being tested or the interpretation of the results (Budd et al., 1999).

Such studies make clear that retracted publications are not being identified appropriately and removed from the literary canon, and that their continued use could contribute to issues of reproducibility. A long-standing concern has been the lack of transparent information on retracted works, and inconsistencies in the format (Pfeifer and Snodgrass, 1990), while there had been no reliable source of such information. Even as the *RetractionWatch* Leaderboard (http://retractionwatch.com/the-retraction-watch-leaderboard/) notes, there are inconsistencies in the number of retractions between different sources, and a Medline search of 233 retracted articles revealed that 22% gave no mention of the retraction (Decullier et al., 2013). While steps to improve identification of retracted articles are improving, it is taking longer to issue retractions (Fang et al., 2012), perhaps due to a lengthy investigative process, after which it can take up to 3 years for retraction notices to appear on *PubMed* (Decullier et al., 2014).

It is particularly troubling when citation of retracted studies is used to justify clinical studies and place patients in harm's way. Evaluating 180 retracted clinical papers, Steen (2011c) found 851 citations that induced prospective clinical trials treating patients. While over 28,000 patients were enrolled and 9,189 treated in studies directly related to the retractions, some 400,000 and 70,501 patients were enrolled and treated, respectively, in secondary studies "that drew ideas or inspiration from the primary study." While some patients were likely harmed, it is difficult to discern how many patients were compromised after the original study was retracted, since the design and execution of clinical trials takes significant time. Steen (2011c) cites one particular example where studying the combination of chemoembolization and radiofrequency ablation in patients with liver cancer that was found to be fraudulent (Cheng et al., 2008), still spawned 144 publications on the faulted procedure after the retraction was

published, even if they did not all cite the Cheng article. This demonstrates the insidious nature of retracted studies. They can perpetuate erroneous science placing patients at risk, through secondary continuation of bogus ideas that are hard to trace back to their fraudulent beginnings but become entrenched in the literature.

1.10.3 The Spectrum of Irreproducibility

The spectrum of irreproducibility is the product of behaviors that extend from calculated, often serial, fraud, termed pathological (Pulverer, 2016) at one end to the "slippery slope" of erroneous design and analysis that are the product of ignorance or a misguided or poorly thought through hypothesis at the other (Curtis et al., 2015; Mullane et al., 2015). Whether such behaviors constitute fraud depends on whether the author deliberately did things they knew were wrong. Unfortunately, the degree of ignorance and hubris in the sector means that the incidence of malpractice that is not strictly speaking fraud is probably immense.

1.10.4 Research Misconduct

A metaanalysis of 146 reports of scientific misconduct from the US Federal Office of Research Integrity (ORI) that covered the period 1992–2003, by Kornfeld (2012) concluded that individual research misconduct could be ascribed both to environmental factors—pressure to publish and perverse incentives (Alberts et al., 2014; Ioannidis, 2014) that included academic and/or financial rewards—and individual psychological traits with the latter being divided into the following categories:

- *the desperate*, whose fear of failure overcame a personal code of conduct;
- *the perfectionist*, for whom any failure is a catastrophe;
- *the ethically challenged*, who succumb to temptation;
- *the grandiose*, who believe that their superior judgment did not require verification; and
- *the sociopath*, who is totally absent a conscience.

Scientists both in academia and industry can be sorted into these aforementioned categories, with the caveat that given the multiple interactions and interdependencies of applied research and its deliverables, there are far more checks and balances and peer and management oversight in industry (as compared with academia) that can lead to the rapid identification of research misconduct resulting in more tangible consequences, for example, demotion, reassignment, or termination for cause.

1.10.5 Fraud

Issues with fraud in biomedical research were the subject of US Congressional hearings as far back as 1981 when key witnesses, described as "über establishment" figures in the biomedical research community, generated consternation when they described the problem of scientific fraud as being "exaggerated" and "not a matter of general societal concern." This led the Chairman of the hearings, one Congressman Albert Gore (of climate change, hanging chads and Nobel Prize fame) to express perplexity since the community spokespersons considered "the problem of fraud…. a private one…. that should be dealt with by

informal codes of the scientific community… and was not an important ethical problem.. [to].. worry those…charged with the public trust" (Gross, 2016). While this attitude is still a prominent one evidenced by many researchers in biomedical research (who fiercely defend their perceived independence), and their institutions, it is not shared by Begley, Ioannidis, the anonymous sages at the *Economist*, or two NIH scientists, Walter Stewart and Ned Feder. Stewart and Feder who were harassed by the hierarchy at the NIH for their investigation of "scientific misdeeds" until they ultimately left. The reader is referred to Gross' engaging and insightful accounts of this and examples of scientific misconduct in the 1970s and 1980s for additional detail (Gross, 2016).

In defining scientific fraud as "serious misconduct with intent to deceive…. the very antithesis of ethical behavior in science," Goodstein (1995) presciently noted that because of increasing competition for scarce funding resources, fraud in science would increase. This viewpoint was revisited by Triggle and Miller (2002) and some 2 decades later by Alberts et al. (2014) and documented by Steen et al. (2013) and others. Goodstein further stated that while "Science is self-correcting….[since fraud]...will eventually be discovered and rejected..[this].. does not protect us against fraud, because injecting falsehoods into the body of science is never the purpose of those who perpetuate fraud." The expectation that science continues to be effectively self-correcting and, as a result, makes fraud pointless, illustrates the difficulty facing "meticulous scientists" (Lemaitre, 2016), who describe themselves as "rigorously honest," making it difficult for them to conceptualize why others, especially those individuals documented in the following sections, would engage in serial fraud that destroys their careers.

1.10.6 Notable Examples of Fraud—Biomedical Researchers Behaving Badly

When articles are retracted in the area of biomedical research on the basis of demonstrable fraud, they are often subjected to deep investigation and hindsight. This has the potential to serve as a helpful catalyst for efforts to find practical solutions to rectify the problem (Bartlett, 2015). While there are many instances of fraud documented on the *RetractionWatch* website (http://retractionwatch.com), the following five were selected as landmarks for reasons that will hopefully be self-evident (Table 1.1). In reviewing the literature on fraud in biomedical research, the interested reader will find other egregious examples including that of the cardiologist, Darsee (Kochan and Budd, 1992) who is number 24 in the list of the 30 researchers with the most retractions on the *RetractionWatch* leaderboard with 17 retractions (http://retractionwatch.com/the-retraction-watch-leaderboard/). At the time of his serial frauds, Darsee outraged the scientific community; over 30 years later, his transgressions may be viewed as tame compared with the 23 other researchers above him on the list, a sign of a continuing and progressive lapse in ethical standards? Or the transparency enabled by the Internet?

1.10.6.1 MMR Vaccine: Andrew Wakefield—Royal Free Hospital

Following from an initial paper that proposed a link between the measles virus and Crohn's disease (Wakefield et al., 1993), Wakefield and coworkers subsequently published a paper in *Lancet* in 1998 (Wakefield et al., 1998) suggesting a link between the use of measles, mumps, and rubella (MMR) vaccine and Crohn's disease. The "findings" of this association could not be reproduced and were in time independently refuted by the Department of Health and

TABLE 1.1 Serial Fraudsters

Investigator	Topic	Detail	Societal consequences	Resolution
Wakefield	MMR vaccination causing Crohn's Disease, "autistic enterocolitis, autism"	Irreproducible "bogus data" supporting vaccine-induced autism pathology Conflicts of interest involving litigation against MMR vaccine manufacturers and Newco to develop diagnostics and treatment for "autistic enterocolitis"	Reduction in MMR vaccine use from 92% in 1995 to 84% or less in 2002. Measles outbreaks in England, Wales, and California.	Retraction of original paper. Investigator struck off medical register.
Woo-Suk	Human embryo-derived stem cell lines for personalized therapeutics	Data on 9/11 of cells fabricated with ethical lapses in the collection of human eggs. Fraud, embezzlement of research funds, and bioethics		Suspect papers retracted. Patent issued for NT-1 stem cell line. Investigator dismissed from Seoul National University and now running South Korean Institute making genetically identical copies of dogs for pet owners
Stapel	Human psychology studies	Investigator planned hypotheses, methods, collection, and outcomes of his experiments and then pretended to run the experiments to gather "too good to be true" results. Denied collaborators and students access to fraudulent "raw data."	Led to field of social psychology being viewed as unique in its ability to manipulate/create data to conduct research fraud. "Career-killing behavior" for colleagues and students	58 suspect papers retracted. Investigator dismissed. Questioning of core validity of social psychology research—for example, a science with "fuzzy" endpoints like the "absurd hypothesis that listening to a Beatles song could make you 1.5 years younger". Autobiography—*Ontsporing* (*Derailed*) (2012) reflecting on circumstances and motivation for fraud.

Obokata	Creation of pluripotent STEM cells using in vitro stress manipulation	Within 4 months 133 attempts to replicate in seven labs failed. STAP cells not a genetic match with mice of origin probably "normal" embryonic stem cells investigator plagiarized text and manipulated images in the papers. Extraordinary claims not supported by extraordinary data.	Contribution to suicide of supervisor. Questioning of rigor of peer review at Nature.	Retraction of suspect papers 7 months after publication. Investigator resigned from RIKEN Institute. Authored book *Ano Hi (That Day)* (2016) claiming she was "framed" and implicating mentor in fraud. Set up "STAP HOPE PAGE" website in 2016 with instructions for making STAP cells.
Potti	Microarray analysis of human tumors to derive a drug response signature to predict the patient response to chemotherapy	Forensic bioinformatics assessment identified "careless, inexplicable errors" in Potti findings. Concerns also expressed within Duke by student dismissed as a "difference of opinion."	Microarray drug response signature used to design and conduct clinical trials in 117 patients after concerns raised. *60 Minutes* (March 5, 2012)—"one of the biggest medical research frauds ever—one that deceived dying patients, the best medical journals, and a great university" http://www.cbsnews.com/news/deception-at-duke-fraud-in-cancer-care/.	Investigator resigned from Duke. Resolution of fraud took 7 years. 2016 lawsuit accusing Duke University of engaging in a civil conspiracy compromising clinical trials in cancer patients.

Medical Research Council in the United Kingdom (Siva, 2010) and a Japanese research group (Iizuka et al., 2000). Nonetheless, the unfounded MMR vaccine association was additionally extended to other bowel disorders and to autism with the "discovery" of a new syndrome described as "autistic enterocolitis" (Deer, 2010), an autism that was thought by its discoverers to be worthy of a Nobel Prize (a thought that is unusually high on the list of delusional career aspirations of those engaged in overt scientific fraud). It was also the basis of a speculative lawsuit, the supportive data for which was highly suspect (Deer, 2010). The original Wakefield et al., paper was ultimately retracted in 2010 (Editors of the Lancet, 2010) on the basis of false claims "that children were 'consecutively referred' and that investigations were 'approved' by the local ethics committee." Major conflicts of interest were also identified (Goodlee et al., 2011) that included Wakefield's litigation against MMR vaccine manufacturers (Deer, 2011a) and the founding of a new company to develop diagnostics and treatment for "autistic enterocolitis" (Deer, 2011b). A subsequent retrospective cohort study (Jain et al., 2015) involving 95,727 children some of whom had siblings with what was now termed autism spectrum disorders (ASD) concluded that the "receipt of the MMR vaccine was not associated with increased risk of ASD, regardless of whether older siblings had ASD. These findings indicate no harmful association between MMR vaccine receipt and ASD even among children already at higher risk for ASD."

Despite these refutations of the "findings," the Wakefield paper led to a grass roots anti-vaccination movement that resulted in a reduction in MMR vaccine use, from a level of 92% in 1995 to 84% or less in 2002 reducing herd immunity to a level well below the 90%–95% required to protect the entire population. In 1998, the year that the original autism-association article was published, 56 measles cases were reported in the United Kingdom. A decade later, in 2008, measles had become endemic in England and Wales with 1,348 cases and 2 confirmed deaths (Thomas, 2010; Hiltzik, 2014). A similar measles outbreak in California in early 2015 was also ascribed to a lack of, or incomplete, vaccination (Majumder et al., 2015). Despite the unusual weight of scientific evidence against the MMR-autism association and the fact that Wakefield had been "struck off" as a licensed physician in the United Kingdom, he made a movie entitled "*Vaxxed; From Cover Up to Catastrophe*" that represented his original premise, and was controversially shown at the Tribeca Film Festival in March, 2016 (Hoffman, 2016; Senapathy, 2016).

When a solidly refuted research finding is championed by the uninformed—often celebrities who lack scientific training but who provide an imprimatur of faux credibility to the topic (Tarkan, 2016)—the outcome is often a pernicious cult that is detrimental to public well-being. The fact that the perpetrator is often immune from the criminal prosecution that might be anticipated given the negative impact of the fraud reinforces the belief of acolytes that the perpetrator is right and that the system is wrong despite "the energy, emotion, and money ..[being],.. diverted away from efforts to understand the real causes of autism" (Goodlee et al., 2011) and the associated suffering and death. If they did what they were accused of doing and have not been convicted, how could the accusation have any merit?

1.10.6.2 Embryonic Stem Cells: Hwang Woo-Suk—Seoul National University

A stem cell researcher in South Korea, Hwang published papers in 2004 in which he claimed to have extracted 11 stem cell lines from human embryos that had the potential to be used as personalized therapies (Hwang et al., 2004, 2005). The data on nine of these proved to have

been fabricated with ethical lapses in the collection of the human eggs used (Cyranoski, 2004; Cyranoski and Check, 2005). These papers were eventually retracted with Hwang being charged with fraud, embezzlement of research funds, and bioethics violations. Despite this, he has continued his research in stem cell science and currently heads an institute in South Korea that makes genetically identical copies of dogs for pet owners (Cyranoski, 2014a) based on work conducted concurrently with the retracted stem cell work (Lee et al., 2005). While ethical issues remain, the scientific status of the disputed papers appears to be unresolved with patents being issued after the retractions for at least one of the claimed stem cell lines, NT-1 (Cyranoski, 2014a).

1.10.6.3 Environment and Human Behavior: Diederik Stapel—Tilburg University

A "Wunderkind" social psychologist at Tilburg University in the Netherlands, Stapel currently holds the number 3 position on the *RetractionWatch* Leaderboard with 58 retractions, having fabricated many of his publications in the field of social psychology, for example, human behavior. One such fabrication was a prominent paper in *Science* that suggested that environmental untidiness led to racism (Stapel and Lindenberg, 2011).

Stapel prefabricated entire experiments (Stroebe et al., 2012; Verfaellie and McGwin, 2011) with his *modus operandi* being to construct the hypotheses, methods, data collection, and outcomes of his experiments and then pretend to run the experiments on his own at local schools. He would then fabricate the data and provide these to apparently unsuspecting colleagues and students for further analysis (Jump, 2011; Stroebe et al., 2012). When these individuals requested access to the raw data they were routinely rebuffed. Stapel's fraud was finally revealed by whistleblowers at Tilburg University and led to his suspension, censure, and ultimate dismissal (Stroebe et al., 2012). In an interim report, the University stated that his behavior had caused "severe damage to young people at the beginning of their careers, as well as to the general confidence in science, in particular social psychology" (Levelt Committee, 2011). Many of the scientists being mentored by Stapel evidently completed their theses without ever doing any actual experiments. His colleagues in the field of social psychology underwent a bout of major soul searching as to why his "too good to be true" fraudulent results (Jump, 2011; Stroebe et al., 2012) went undetected for so long (Verfaellie and McGwin, 2011) and eventually ascribed it to "a general culture of careless, selective, and uncritical handling of research and data." (Bhattacharjee, 2013). Some colleagues were also indignant that the final report from Tilburg University read as if Stapel's fraud was a phenomenon unique to the field of social psychology with one individual noting that "there are no grounds for concluding either that research fraud is any more common in social psychology than other disciplines or that its editorial processes are particularly poor at detecting it" (Gibson, 2012; Stroebe et al., 2012). Another colleague noted that "to understand fraud, we should think about how it begins and escalates, not how it ends. By the time such fraud is exposed, bad choices that would usually lead to only minor transgressions have escalated into outright career-killing behaviour" (Crocker, 2011).

In a follow-up article entitled *The Mind of a Con Man* that involved extensive interviews with Stapel, Bhattacharjee (2013) noted that "Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct

sexy results that journals found attractive." "It was a quest for aesthetics, for beauty—instead of the truth." While Stapel's misdeeds have, due to their magnitude, tended to occupy center stage in recent reports of fraud in the social psychology sciences, others including Dirk Smeesters, Lawrence Senna (Yong, 2012), and Marc Hauser (Wade, 2010), have been similarly suspected or proven responsible for manipulating data, both human and animal. This has led to widely-held concerns that psychology, because of its reliance on self-reporting from subjects, is a discipline with soft, "fuzzy" endpoints. This viewpoint is reinforced by outcomes in social psychology like those reported in a study in which evidence was generated (Simmons et al., 2011) for the "absurd hypothesis that listening to a Beatles song could make you 1.5 years younger" (Estes, 2012).

A take home from this example of egregious fraud, however is not nuanced. If it is easy to get away with fraud in a particular discipline because the practitioners of that particular discipline are poorly trained in the scientific method, gullible, complacently uncritical, and naïve, this does not excuse the fraud. Fraud is fraud. The issue for the discipline, however, is how to acquire and implement the necessary vigilance, know-how and expertize, as exemplified by rigorous reporting requirements and thorough peer review, that would ensure that fraud cannot walk through the front door of subject publications, unmolested, in the future, while not destroying the core aims of scientific inquiry (Woodward and Goodstein, 1996).

1.10.6.4 Stimulus-Triggered Acquisition of Pluripotency (STAP): Haruko Obokata—RIKEN Institute

Two papers published in *Nature* reporting the reprogramming of mammalian somatic cells using a stressor, for example, transient low-pH—citric acid, to generate pluripotent stem cells (Obokata et al., 2014a,b), were viewed as a major breakthrough in the "hot" field of stem cell research in January of 2014 (De Los Angeles et al., 2015; Goodyear, 2016; Normile and Vogel, 2014; Rasko and Power, 2015). However, by April 2014 as the result of numerous failed attempts—some 133 in 7 laboratories—to reproduce these findings, RIKEN the host institution, found Obokata guilty of research misconduct. By July 2014 these papers were retracted.

Of especial note in this instance of fraud was the short time period, some 6 months between the publication of these papers and their retraction that no doubt reflected the potential importance of the finding and an increased awareness and concern regarding the importance of the failed replication. There was also a positive side in that the detection of the fraud was a welcome instance of the scientific community successfully self-correcting.

Like Stapel, Obokata was considered a "star" both in her area of stem cell research and as a symbol of Japan's *rikejo* ("science women"; Hongo, 2014) with inevitable rumors of a Nobel Prize for her breakthrough STAP research. When the fraud was uncovered and the outcomes reported at the now obligatory forum of a press conference, Obokata "apologized for many things that day. She apologized for insufficient efforts, ill-preparedness, and unskillfulness, for errors of methodology, and sloppy data management. They were all, she said, benevolent mistakes, due to her youth and inexperience. But she denied fabricating her results" (Rasko and Power, 2015). It was also found that the STAP cells did not genetically match the mice from which they originated indicating they were probably "normal" embryonic stem cells and that Obokata had plagiarized text and manipulated images in

the papers. As final closure, when Obokata was provided the opportunity by RIKEN to repeat her own experiments, she was unable to do so (Hongo, 2014). Nonetheless, in 2016, Obokata launched a website "STAP HOPE PAGE" (Jiji, 2016) to help researchers reproduce her findings.

In a more personal context, her case raised issues as to whether aggressive reporting in the scientific and mainstream media violated her human rights (Stemwedel, 2015) and whether the latter was a factor in the suicide of her supervisor, Yoshiki Sasai (Cyranoski, 2014b).

An additional wrinkle in the STAP saga was that one of the coauthors of the original papers (Obokata et al., 2014a,b) was Charles Vacanti, at one time the Chairman of Anesthesiology at Brigham and Women's Hospital in Boston who claimed that he and his brother had originated the STAP concept (Goodyear, 2016). Obokata had conducted postgraduate work in Vacanti's laboratory before joining RIKEN and even after the STAP fraud issues emerged, Vacanti still claimed, without providing any evidence (Goodyear, 2016; Rasko and Power, 2015), that he could create STAP cells, this despite the continued and consistent failure of others to do so (De Los Angeles et al., 2015).

1.10.6.5 Microarray Genetic Analysis for Personalized Cancer Treatment: Anil Potti—Duke University Medical Center

In 2006 Potti, together with his supervisor Joseph Nevins, published a paper in *Nature Medicine* (Potti et al., 2006) claiming that a microarray analysis of human tumors could be used to derive a drug response signature that could predict the individual patient response to chemotherapy. The discovery of a reliable "omics-based predictive diagnostic was important in that it provided a means to design individual patient treatment and thus avoid the trial and error approach common with treating cancer." Because of the importance of these findings and the failure of previous predictive tests, two biostatisticians at the MD Anderson Cancer Center, Baggerly and Coombes, used forensic bioinformatics to assess the Potti findings and found errors "some….careless…[and]..others… inexplicable" (Kolata, 2011). These concerns were initially published in a note in *Nature* (Coombes et al., 2007) and later in greater detail in the *Annals of Applied Statistics* (Baggerly and Coombes, 2009). Despite the initial *Nature* note, the Potti findings were used in 2008 as the basis to design and conduct clinical trials in 117 cancer patients. These trials were suspended in 2009 following the publication of the full Baggerly and Coombes paper that led to the retraction of several high-profile papers by Potti. The concurrent identification of overt misrepresentations in Potti's CV together with 11 cases of malpractice (http://retractionwatch.com/2015/05/01/malpractice-case-against-duke-anil-potti-settled/) also resulted in his resignation from Duke. A patient lawsuit accusing Duke University of engaging in a civil conspiracy (Goldberg, 2015) was reportedly resolved (Ramkumar, 2015).

Internal concerns regarding Potti's research were documented in an email in March, 2008 by a concerned medical student in the Nevins laboratory, Bradford Perez. This was downplayed by the Duke Administration as a "difference of opinion" and only came to light as a result of an in-depth report from the *Cancer Letter* in 2015 (Goldberg, 2015). Unlike the rapid resolution of the Obokata STAP fraud, that conducted by Potti took 7 years to resolve (http://retractionwatch.com/2015/11/07/its-official-anil-potti-faked-data-say-feds/). Additional detail on the Potti case is available in the NASEM (2017) report pp. 234–240.

1.10.7 Fraud as an Opportunity for Reflection and a Learning Moment?

Of the five egregious examples discussed (Table 1.1.), the retractions in the preclinical research conducted by Stapel, Hwang, and Obokata have, at least in the short term, had little direct impact on patient care although they have had a major impact on research directions in the areas of cloning, social psychology, and stem cell research. Conversely, the papers published by Wakefield and by Potti directly affected the practice of medicine, in the former instance leading to the establishment of an antivaccination movement, the actions of which have put others at unnecessary risk and led to a controversial mandatory vaccination law in California (Salmon et al., 2015) while the latter compromised clinical trials in cancer patients.

It is also noteworthy that the number of papers that are ultimately retracted is immaterial, with the single case of Wakefield doing more harm to patients and society than the 183 that were retracted in the case of Fujii.

There is a natural inclination to downplay the problem of fraud by ignoring it or treating it as an aberration. A far better reaction is to assess whether there are lessons to be learnt from the fraudsters themselves and the reactions of their institutions and funding organizations, lessons that will aid in tangibly improving the reproducibility, fidelity, quality, and relevance of biomedical research. This is certainly the viewpoint that was captured under the rubric "Can a Longtime Fraud Help Fix Science?" in the blog by Bartlett (2015) and was shared by a number of authors including Kakluk (2009), Borsboom and Wagenmakers (2013), Witkoski (2014), and Berry (quoted in Goldberg, 2015).

For additional insights in understanding and dealing with scientific misconduct and fraud in biomedical research together with recommendations and best practices to address these issues, the reader is referred to the monograph, *The Management of Scientific Integrity within Academic Medical Centers* (Snyder et al., 2015).

1.11 REPRODUCIBILITY AND TRANSLATIONAL MEDICINE

More than half of the failures in Phase II clinical trials of new drug candidates are due to lack of efficacy (Harrison, 2016; Kimmelman and Federico, 2017). Concerns regarding the reproducibility of animal findings are routinely cited as one contributing factor in the poor translation of research findings to the clinical setting where translation is described as "the transfer of new understandings of disease mechanisms gained in the laboratory into the development of new methods for diagnosis, therapy, and prevention and their first testing in humans" (Sung et al., 2003).

So while reproducibility and translation are not synonymous or interchangeable concepts (Jarvis and Williams, 2016), they do overlap. The issues that influence reproducibility—large effect sizes in underpowered studies, inadequate experimental design and analytical methods, use of poorly validated tools and models, etc.—also impact effective translation as they undermine the robustness of the observation, a metric that goes beyond reproducibility. Consequently, even if a finding is reproducible, if it is not of a magnitude that can be meaningful when applied to a clinical setting, and observed in a population with diverse genetic and epigenetic backgrounds, then it has limited utility. A second departure for translatability being differentiated from reproducibility is around the endpoints selected for demonstrating

efficacy. In the research setting, it is relatively easy to incorporate either mechanism-related measures to assess compound activity, for example, collagen deposition in a pulmonary fibrosis model, as a proxy for clinically applicable endpoints of lung function, such as forced vital capacity, or to rely on surrogate measures of efficacy—for example, contextual fear conditioning (Comery et al., 2005) or a Morris water maze test (Jian et al., 2016) in transgenic models of Alzheimer's disease—that bear no relationship to the standard clinical measures, such as the Clinical Dementia Rating scale (CDR sum of boxes; O'Bryant et al., 2008) that addresses six functional domains related to memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. Generally the argument is made that there is a correlation or temporal association between the pathology and the measures used in animal models, although such associations alone constitute scientifically weak evidence, as demonstrated in numerous amusing examples on the Internet between per capita consumption of margarine in the United States and the divorce rate in Maine; automobile drivers killed in a collision with a train and US crude oil imports from Norway (http://www.tylervigen.com/spurious-correlations); and *iPhone* sales and individuals dying from falling down stairs (https://hbr.org/2015/06/beware-spurious-correlations). However, more important to this thesis is that even if such studies were reproducible, their translatability to the clinical setting, based on dubious associations, constitutes a major leap of faith.

Predicated on a Phase I clinical trial of the fatty acid amide hydrolase (FAAH)-inhibitor BIA 10-2474 that led to serious neurological complications including one fatality, Kimmelman and Federico (2017) argued that "a lack of emphasis on evidence for the efficacy of drug candidates is all too common in decisions about whether an experimental medicine can be tested in humans" and recommended that the regulatory bodies—including the institutional review boards (IRBs) responsible for approving clinical trial protocols place an emphasis equal to that on compound safety on understanding compound efficacy. They propose "three questions to assess clinical promise":

1. *What is the likelihood that the drug will prove clinically useful?*
   - How have other drugs in the same class or against the same target performed in human trials?
   - How have other drugs addressing the same disease process fared?
2. *Assume the drug works in humans. What is the likelihood of observing the preclinical results?*
   - Are the treatment effects seen in animals large and consistent enough to suggest a tangible benefit to patients?
   - How well do animal models reflect human disease?
3. *Assume the drug does not work in humans. What is the likelihood of observing the preclinical results?*
   - Have effects of random variation and bias been minimized (e.g., by sample sizes, randomization, blinding, dose-response curves, and proper controls)?
   - Do the conditions of the experiment (for instance age of animal models, timing of treatments, and outcomes) match clinical scenarios?
   - Have effects been reproduced in different models and/or in independent laboratories?

Providing answers to these sorts of questions is far from a trivial undertaking and has been likened to "translat..[ing].. a text with the sophistication and depth of Shakespeare using a first-grader's vocabulary and experience, because our knowledge about the functions of most

pathways in various cell types, during different developmental stages, and under normal physiological conditions, is still rudimentary and piecemeal" (Zoghabi, 2013).

Nonetheless, such considerations underline the need for both ensuring that the preclinical data used as the basis of moving a new drug candidate into clinical trials be "true" as defined by Goodman et al. (2016), for example, reproducible and generalizable, and also have context in relationship to the reported effects of other known compounds, if any, acting at the same disease target (Kimmelman and Federico, 2017).

1.11.1 Animal Models—Predictive Value in Translational Research

A key—and often the penultimate—step in the preclinical drug discovery process is to demonstrate the efficacy of a drug candidate in an animal model that is thought to reflect aspects of the human disease state that consequently has predictive validity (Drucker, 2016; Groenink et al., 2015; McGonigle and Ruggeri, 2014; NAP, 2015; van der Worp et al., 2010). For example, an animal model of pain where morphine, acting directly via morphine receptors, has been repeatedly shown to reverse pain behavior induced by chemical or physical trauma may be considered as predictive for the testing of new compounds that have the potential to work in similar human pain states. Unfortunately, in this example, the translational models have proven to be far less predictive than preclinical data would suggest, only being valid for compounds that act via an opioid receptor. Newer analgesics that act via one of the many (more than 20) novel mechanisms identified in the past 20 years that show robust and reproducible efficacy in a morphine model of pain yet have consistently failed to produce analgesia in humans, indicates the high level of false positives in preclinical studies (Fairbanks and Goracke-Postle, 2015).

Animal models of human disease states remains a contentious issue throughout the spectrum of preclinical research (Denayer et al., 2014; Greek and Hansen, 2013; Peters et al., 2015; Sams-Dodd, 2006; Vandamme, 2015) with many researchers confusing the inability to predict the translation of the therapeutic activity of NCEs observed in animal models to the clinic with failed reproduction *per se*. Indeed, many NCEs that evidence robust, dose-dependent and replicated activity in preclinical models often fail in the clinic. As noted by Grove (2005) "Humans are incredibly complex biological systems, and working with them has to be subject to safety, legal, and ethical concerns.... The result is wide-scale experimentation with animal models of dubious relevance, whose merit principally lies in their short lifespan."

The fidelity of animal models of disease to the human condition is generally low, irrespective of the therapeutic area, given species differences in biological mechanisms and systems pharmacology between rodents and humans, and also because the mechanistic causality of many diseases remains unknown (Groenink et al., 2015; Llovera et al., 2015). As a result, many animal models, especially transgenics, are exclusively models of presumed and often unproven molecular mechanisms of a disease rather than disease states per se. Additional shortcomings of animal models relate to the putative disease phenotype in the animal which is temporally more acute than that of the actual human disease state. Thus chronic pain in humans reflects a subtle and complex process that may be decades in development with extensive neuronal remodeling. This is in stark contrast with chemically—or surgically-induced animal models of pain that are acute (hours) to subchronic (weeks) in onset and duration and focus on a limited repertoire of known or putative pain-associated targets.

Similarly, Type 2 diabetes (T2D) research is generally modeled in C57BL/6J mice 2–6 months of age as the animals are readily available and cost effective (Drucker, 2016). However, in humans T2D onset occurs in the fifth through ninth decades of life by which time humans have suffered years of low grade tissue inflammation, hypertension, dyslipidemia, fibrosis, weight gain, and impaired glucose tolerance. Since younger animals have greater organ and cellular plasticity and cell proliferation, they present a background that generally makes it easier for putative therapeutics to reverse a disease phenotype than that present in older animals, thus confounding their use in translational research (Drucker, 2016). Furthermore, in genetic models, the knockout or overexpression of a discrete disease-associated target is an artificial, reductionistic monogenic approach with many, if not the majority, of human diseases being multifactorial involving contributions from dozens of genes that have modest effect sizes (Fuchsberger et al., 2016). Gene effects on disease onset often occur in early stages of development and also involve environmental contributions. For instance, in schizophrenia, a disease that involves alterations in early brain development (Kahn et al., 2015), researchers have been looking for a single susceptibility gene (the one disease, one gene mantra) for more than 2 decades. As of 2008, schizophrenia has been associated with 3,608 polymorphisms in 516 different genes (Allen et al., 2008) suggesting that the underlying genetics of the disease are complex and that identifying a major susceptibility gene is unlikely (Farrell et al., 2015). An additional confound of using germline knockouts as translational disease models is their potential "for developmental adaptation and physiological compensation secondary to the loss of the key gene and protein at the earliest stages of development" (Drucker, 2016). Adaptation and compensation can lead to systems redundancy (Doyle et al., 2012) where a gene knockout can be compensated for by an alternative pathway. Thus, preclinical efficacy that supports compound advancement to the clinic can frequently be replicated to an extreme level of precision and still be reflective of a model that is not predictive for the human condition. In fact, the absence of efficacy in the clinic, the lack of translation of an NCE (New Chemical Entity), has become, like that of the preclinical animal models, very reproducible, as discussed in the examples below.

1.11.2 Limitations in Animal Models of Human Disease

The lack of effective therapeutics to treat chronic disease states, for example, AD, T1D, and neuropathic pain, has prompted considerable efforts in basic research to meet "unmet medical needs." Despite this imperative, many potential therapeutics with convincing drug-like properties and appropriate efficacy in animal models that are thought to recapitulate molecular and phenotypic aspects of the human disease have consistently failed in the clinic—prompting Greek and Hansen (2013) to note that in "reductionism-based animal models, the failures have been epic."

1.11.2.1 Amyotrophic Lateral Sclerosis (ALS)

A number of compounds reported in approximately 50 papers published before 2008 were identified as having survival benefit in the standard ALS $SOD1^{G93A}$ mouse model but failed to translate to the clinic and provide human benefit. This led to the conclusion that "the high noise floor of the model and the failure of the selected studies to replicate support the conclusion that the bulk of published studies using the $SOD1^{G93A}$ mouse model may unfortunately be

measurements of biological variability due to inappropriate study design" (Scott et al., 2008). Blinding and randomization were conspicuously lacking in this example.

Additional testing of a cohort of approximately 100 potential therapeutics that had been reported preclinically to attenuate ALS symptoms in animal models found that the original preclinical findings could not be replicated (Perrin, 2014). Only 8 of these compounds were advanced to clinical trials, but subsequently failed to show efficacy indicating that the preclinical data used for their advance to the clinic was a reflection of false positive findings. This was attributable to limitations in the ALS mouse model or the design and analysis of studies that used it. Specifically, in ALS patients and some mouse models of the disease, the paralysis caused by deterioration of the neurons innervating skeletal muscle progresses over time. However, in the ALS mouse model where a mutant form of the RNA-binding protein TDP43 was associated with motor neuron loss, protein aggregation, progressive muscle atrophy, and a defective version of the SOD1 gene (which is mutated in 10% of the familial ALS population), this progressive deterioration was not observed (Perrin, 2014). Additionally, muscle deficits in TDP43 mice were mild with animals dying of acute bowel obstruction caused by deterioration of gut smooth muscle rather than the progressive muscle atrophy observed in human ALS (Hatzipetros et al., 2014).

Further analysis of the failure to replicate the original findings identified four factors that increased variability (Perrin, 2014) and that reflected a lack of user consideration of the nuances in using animal models of ALS: (1) a failure to exclude animals whose deaths were unrelated to the disease; (2) not randomizing littermates between control and treatment groups; (3) not accounting for gender since male mice demonstrate ALS symptoms as much as a week before females and die approximately a week earlier—differences that could be construed as a compound-related effect; and (4) the loss of disease phenotype due to multiple copies of the disease-causing gene building up with breeding in a mouse colony that were not passed on in a stable fashion as cells divide leading to subsequent generations having fewer copies of the transgene and a less severe disease phenotype.

More recent studies have focused on TDP-43 misfolding rather than SOD1 dysfunction as a causative factor in ALS. This has highlighted the benefits of attenuating the loss of function and gain of function/dominant-negative toxicity role of TDP-43 in initiating ALS by preventing its misfolding and/or enhancing its clearance, the latter using enhancers of chaperone-dependent TDP-43 folding, and activators of the ubiquitin proteasome system and autophagic pathways (Scotter et al., 2015). While intellectually appealing, it remains to be seen if this concept has any greater success at clinical translatability.

1.11.2.2 Stroke

Considerable efforts have been expended over the past 4 decades to identify effective interventions to reduce mortality and improve outcomes from stroke. These include a variety of small molecule therapeutics acting through targets thought to be directly involved in mediating the excitotoxicity and free radical formation that occurs subsequent to an ischemic event (van der Worp and van Gijn, 2007).

Despite a huge investment of resources, of more than 1,000 putative neuroprotective NCE treatments demonstrating benefit in animal models [mainly in the MCAO (middle cerebral artery occlusion) model in gerbil or rat], 114 of which were examined in clinical protocols in which aspirin and thrombolytic (e.g., alteplase, rTPA) had shown robust efficacy, none were

efficacious (O'Collins et al., 2006). Subsequent analysis of these results (Macleod et al., 2009) identified several variables, including timing of NCE administration as well as animal age, comorbidities, and physiological status, as contributing to the disparity between findings from the animal models and the clinical trial outcomes reflecting bias in the preclinical models that resulted in an "overstatement of neuroprotective efficacy" (Sena et al., 2010). That these models are still being used in preclinical stroke research—more than a decade after extensive peer-reviewed metaanalyses of their total lack of translational value was published—attests to laziness or hubris.

1.11.2.2.1 METAANALYSES OF DATA FROM ANIMAL MODELS OF STROKE

In a metaanalysis of animal data (Bath et al., 2009), the clinical failure of the free radical scavenger, NXY-059, in large clinical trials (5,028 patients) in acute ischemic stroke (Diener et al., 2008) was evaluated in the context of the positive preclinical animal model data (Bath et al., 2009). This included reduced infarct volume and motor impairment in experimental stroke models (transient, permanent, and thrombotic) in rodents, rabbits, and primates (Macrae, 2011). Analysis of the data from 585 animals (NXY-treated 332; control, 253) from mice, rats, and marmosets that originated from 12 laboratories which reflected 26 experiments (four of which were unpublished) showed that NXY-059 was neuroprotective in preclinical models that met the established STAIR (Stroke Therapy Academic Industry Roundtable) criteria (Macrae, 2011). There was evidence however, that publication biases were present in the preclinical studies reviewed. Additionally, while spontaneously hypertensive rat (SHR) models of stroke were included in the metaanalysis, NXY-059 was retrospectively found to be effective only in normotensive rats. Another concern was that sample size calculations were absent from all the studies. Thus the discrepancy between the preclinical and clinical data may have resulted from: (1) a lack of relevance of the preclinical data to the human situation; (2) efficacious doses in rats and marmosets not being predictive of the human situation and; (3) issues with brain access of the free radical scavenger.

The metaanalysis concluded that because of the various biases, the preclinical efficacy of NXY-059 may have been overestimated (Bath et al., 2009; Dirnagl and Macleod, 2009; Macleod et al., 2008, 2009). Based on these conclusions, the authors recommended that metaanalysis of all available preclinical data on an NCE be conducted before the initiation of clinical trials.

Additional factors identified in the translational failures in stroke included potential differences between human brain and that of rodents and the initial assessment of novel therapeutics in a stringently controlled preclinical testing environment where the animals used and the laboratory conditions differ markedly from the heterogeneity seen in stroke patients, and the time of compound administration *after* the ischemic episode.

To further address these issues and those raised in the NXY-059 metaanalysis study, a primate (cynomologous macaque) embolic stroke model was used in a blinded crossover study to assess a novel neuroprotectant, Tat-NR2B9c (aka NA-1) which blocks neurotoxic signaling events by uncoupling the postsynaptic protein, PSD-95 (Cook et al., 2012). Primates treated with Tat-NR2B9c *after* the onset of embolic strokes showed reduced stroke numbers and stroke volumes that anticipated the outcomes of the corresponding human trial, ENACT (Evaluating Neuroprotection in Aneurysm Coiling Therapy; Hill et al., 2012), an apparent success in terms of translation.

Another approach to improve the translational process in stroke involved the use of a *preclinical* randomized controlled multicenter trial (pRCT), the design and rigor of which was based on that of a typical Phase III RCT. This study was conducted in six independent research centers using two models of stroke, the cMCAO (permanent distal middle cerebral artery) and fMCAO (transient middle cerebral artery occlusion) in C57BL/6J mice to assess the effects of a CD49d antibody that inhibits leukocyte migration into the brain (Llovera et al., 2015). Standardization of procedures between the six centers followed similar procedures to those used by Crabbe et al. (1999) (Section 1.4.1.1) and involved 315 mice, 81 in the cMCAO model and 174 in the fMCAO model.

Anti-CD49d, given 3 h after stroke induction, consistently decreased leukocyte migration and infarct volume in the cMCAO model that involved small cortical infarcts but not in the fMCAO model where larger lesions were induced. Anti-CD49d had no effect on behavioral outcomes (rotarod test, adhesion removal test) that were confounded by great intercenter variability. These outcomes were concluded to reflect a variety of subtle differences between the two models including: insufficient statistical power in the fMCAO model due to unexpected variability that could have been addressed by increasing group sizes; differences in the time point of assessment due to higher mortality rates in the fMCAO mice; differences in the neuroinflammatory markers between the two models (that were reflected in a twofold higher infiltration of leukocytes in the cMACO versus the fMCAO model); the use of unequal numbers of animals in the two models; an absence of primary data due to mortality or a lack of infarct demarcation; and low performance in the behavioral outcomes. While a major *tour de force* in using a pRCT approach, the Llovera et al. (2015) study still highlighted many of the limitations in rodent animal studies of stroke and their utility in translation. Furthermore, a completed Phase II clinical trial of the anti-CD49d antibody showed no benefits (Clinicaltrials.gov identifier NCT01955707).

1.11.3 Issues in Translatability

In the disease states discussed earlier, all of which are neurological, the ability to predict efficacy in humans based on animal models remains poor (Hartung, 2013; Hobin et al., 2012; Wendler and Wehling, 2012; Wehling, 2009). Attempts to understand the disconnects between preclinical data and the clinical trial outcomes reveal a variety of causes. Some of these, like experimental design and powering and data analysis, and reagent validity are addressed in this monograph. Others remain unknown (Mullane and Williams, 2015) and are therefore insurmountable which makes putative animal models of human disease states less predictive in assessing potential human efficacy.

This is not unique to the field of neurology as numerous examples from other therapeutic areas have questioned the relevance and predictivity of animal models (Groenink et al., 2015; McGonigle and Ruggeri, 2014; NAP, 2015) other than these providing another measure of a pharmacodynamic response to an NCE that is accompanied by a pharmacokinetic component that can aid in dose selection for human testing (Caldwell et al., 2004; Kleiman and Ehlers, 2016).

Vatner (2016) has argued that the shift in research funding from traditional physiology to molecular medicine that has been ongoing for the past 30 years (Jobe et al., 1994) has resulted in a dearth of laboratories with the appropriate physiological expertize, in this particular instance, cardiovascular physiology, along with the many decades of practical experience

associated with the discipline. This has deprived biomedical research of a critical aspect, that of integrating molecular with whole animal data, that has confounded the integrative hierarchy of experimentation (Kenakin et al., 2014) and led to the whole animal aspects of physiology being poorly understood and inaccurately represented. As a result, integrative physiology is treated as an incidental to molecular findings (or to paraphrase a one-time candidate for the Nobel Prize in Physiology or Medicine—"compound A binds to target Z in vitro, let's start clinical trials") that has been replaced with "wishful thinking" (Vatner, 2016). Similar concerns regarding the decline of integrative biology had been made over 20 years ago by Jobe et al. (1994), who questioned the relevance of in vitro studies to physiological mechanisms that probably has been a major contributor to the translational shortcomings of the animal models discussed earlier—reducing them to in vivo test tubes via transgenic manipulations.

1.12 CONCLUSIONS

In setting the stage for the remainder of this monograph on reproducibility, the present Chapter has focused primarily on the seminal commentaries, papers, and events that have driven the debate on reproducibility to its current level of visibility, active debate, and attempted resolution. This has necessarily required a focus on fraud in the biomedical sciences—a topic that is often far more accessible and interesting than, for example, the use of Bayesian statistics, cell line authentication, or translational confounds.

Fraud per se is probably a very minor contributor (estimated by various sources as being 0.02% or less of all research activities) to the overall issue of reproducibility in biomedical research and its resolution is more likely to occur via the actions of grant review bodies, institutional standards, and ultimately law enforcement than peer review as these can wield (but frequently do not) the big stick of tangible personal consequences. However, by its very nature, published fraud and related misdemeanors have received increased attention by scientists invested in "research on research" who have found a ready outlet for their considerable body of work in the open access literature. It has also been gleefully blown out of proportion—as much that occurs in the 21st century increasingly tends to be—by the mainstream media avidly seeking content, the more sensational, the better. Indeed, Drucker, (2016) has noted "The media itself has an extraordinary appetite for scientific and medical information, especially stories with a hint of therapeutic relevance. The media beast is insatiable, although even my mother has now learned that most "medical breakthrough stories" featured on the television, radio, in print, or disseminated via the Internet and social media are almost always exaggerated and often frankly incorrect."

In the majority of instances, articles on "research on research" in the form of metaanalyses and also via blogs, where issues on the validity of published research findings are raised, perform an invaluable service that grant review bodies and research institutions fail to acknowledge. They do however give the inappropriate impression that reproducibility is in crisis, and has recently escalated dramatically, when it is in fact a chronic problem made more visible by the increased transparency offered by the Internet.

To document the occurrence of egregious fraud in a monograph that is intended to address the topic of reproducibility may seem to some as unnecessary as it tends to highlight the most negative aspects of the reproducibility issue. However, as noted previously, serial fraud is

probably only the tip of the reproducibility iceberg and there needs to be a concerted effort on the part of the biomedical research community as a whole to acknowledge fraud in order to effect meaningful change rather than delegating the responsibility for remedying the situation to journal editors and peer reviewers.

Many scientists consider biomedical research fraud as being at one end—the extreme end—of the reproducibility spectrum with its practitioners, like the five discussed previously as not being worthy of being called a scientist. At the other end of the reproducibility spectrum is the phenomenon of an "honest mistake" where ethical and competent scientists can sometimes, inadvertently, generate bad science. Between these extremes are various forms of investigator bias, overt and unintended that can lead to honest mistakes. Collectively, this leads to a perception by many researchers that 99% of his or her colleagues will subscribe to the same scientific values that they were taught and espouse and would, accordingly, express absolute disbelief (as apparently occurred repeatedly with Stapel and Potti) if they were accused of being involved, intentionally or circumstantially (laboratory members making up or manipulating data) in aspects of fraud (Wilson, 2016).

Like the reproducibility spectrum, fraud is also a semantic continuum that ranges from the egregiously fictitious activities of a Stapel to the selection bias evidenced by one of Begley's unnamed colleagues—no doubt a distinguished cancer researcher—who made the decision to ignore five negative findings in a series of six cancer experiments and instead focus on, and publish, the one finding that was deemed to be "the best" (Begley, 2012). This emulated in intent the behaviors of Stapel, Obokata, and others whose sociopathic priorities, situational and financial, demonstrate a lack of personal responsibility that adds to the erosion of public trust in biomedical research (Alberts et al., 2014; Begley and Ioannidis, 2015) and its integrity (Kaiser, 2014). In the bigger picture, the frauds perpetrated by Wakefield, Hwang, Stapel, Obokata, and Potti can be viewed as examples of extreme personal bias, in the same category as the bias that researchers deal with on a daily basis that differs only in degree from that of overt fraudsters.

While there are grey areas in reproducibility, especially where unintentional bias is involved, biomedical researchers can take simple steps to remove ambiguity in their work and ensure that the research conducted in their laboratories is transparent, relevant, and reproducible. Other unintentional biases unique to modern day science might include the assembly and interrogation of large databases where concerns of keeping them updated, accurate, and well-curated requires significant, long-term investments, while continually modified and refined analytical tools can constantly adjust interpretations of the large datasets, but rarely are the preceding publications revised. The research community needs to "grasp the nettle" if they want to ensure that the continued funding of research remains a high priority for the public and lawmakers (Alberts et al., 2014; Moher et al., 2016). This monograph is intended to aid in that endeavor.

In the remainder of this monograph, the various reproducibility issues are documented together with suggested solutions to improve shortcomings. This includes the provision of substantive guidance on aspects of the design, execution, analysis, and reporting of research findings and the separate but important issue of transparency—if it can't be *seen*, how can anyone know it has been *done*?

In addition, consideration is given to the various mechanisms intended to encourage, evaluate, and judge whether guidance to facilitate reproducibility have been followed. These include the following: (1) the more stringent application, oversight, and proactive support of the prepublication *CPR* peer review process (Chapter 5.2) by supplementing, this with Internet-enabled PPPR (Chapter 5.3.2.3) and *PPC* (Chapter 5.2.3.4) which have the potential, if

used correctly, to reinforce the scope, standards, and execution of the process of self-correction (Alberts et al., 2015; Nature, 2013); (2) the use of publication/peer review checklists in concert with more rigorous and detailed guidelines on the requirements of journals for publication (Begley, 2013; Collins and Tabak, 2014; Curtis et al., 2015; Mullane et al., 2015) and; (3) formal reproducibility initiatives (Errington et al., 2014; OSC, 2015; Van Noorden, 2014), the value of which has been questioned (Mullane and Williams, 2017).

Equally important are initiatives focused on improved training and mentoring, represent the best hope for improving reproducibility and include: (1) improving the general standards for the training and mentoring of new generations of scientists (Collins and Tabak, 2014; Flier, 2017; Kornfeld and Titus, 2016; Michael, 2015; Munafò et al., 2017); (2) increasing the focus on logical and transparent design, powering, execution, and analysis of experiments (Chapters 2 and 3); and (3) using properly validated reagents, cell lines, and animal models (Freedman and Inglese, 2014; Freedman et al., 2015; Chapter 2).

The dedicated fraudster will hopefully find little useful in this monograph—other than the possible opportunity to learn about the more recently identified examples of fraud and misrepresentation and, by default, understand how to enhance their efforts in avoiding detection. These individuals certainly will not be deterred by any strategies proposed to diminish their impact on advancing science as opposed to their careers. Instead this monograph is intended as an aid for the other 95 + percent of well-intentioned scientists, who wish to avoid the pitfalls and traps that can ensnare their best efforts, and be able to judge the quality, and hence the reliability and truthfulness of published biomedical research.

1.12.1 A Note on Estimates and Surveys

Throughout this monograph, the authors cite various estimates in the literature and mainstream media of the occurrence of various issues related to reproducibility. These generally range from 0.01 (or less) to 50% and are reported with little in the way of convincing substantiation or authoritative sourcing. This may be due, in part, to the lack of transparency in the process of identifying and proving fraud, the extensive delays in the investigative processes and then further delays in informing journal editors and retractions making their way to *PubMed* or other databases. Also, as mentioned, there is a decidedly grey area between outright pathological fraud and the practices of "sloppy science" with selective reporting, Harking, *p*-hacking, and overt bias in selection, interpretation, and presentation that could be argued as fraudulent practices although the perpetrators clearly would not regard themselves as practicing anything other than the responsible conduct of research (NASEM, 2017). The use of the cited range of anywhere between 0.01%–50% is therefore used with the recognition that while it is broad, it probably underestimates the extent of the problem.

Similarly, several surveys are cited that need to be taken with a grain of salt since they have extremely low response rates that are compounded by the responders self-selecting. For instance, the *Nature* survey on whether there is a reproducibility crisis (Baker, 2016b) that involved 1,576 researchers was, depending on how many biomedical researchers there are in the world, woefully underpowered but nonetheless extensively cited. In the United States, this population is estimated at 70,000–80,000 (Heggeness et al., 2017), while a conservative extrapolation would estimate a global group size as 2–3 times larger making the *Nature* survey group of 1,576 representative of somewhere between 0.66% and 2.25% of the actual biomedical research workforce and thus scientifically questionable.

References

Alberts, B., Kirschner, M.W., Tilghman, S., Varmus, H., 2014. Rescuing US biomedical research from its systemic flaws. Proc. Natl. Acad. Sci. USA 111, 5573–5777.

Alberts, B., Cicerone, R.J., Fienburg, S.F., Kamb, A., McNutt, M., Nerem, R.M., et al., 2015. Self-correction in science at work. Science 348, 1420.

Allen, N.C., Bagade, S., McQueen, M.B., Ioannidis, J.P.A., Kavvoura, F.K., Khoury, M.J., et al., 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat. Genet. 40, 827–834.

Almeida, J.L., Cole, K.D., Plant, A.L., 2016. Standards for cell line authentication and beyond. PLoS Biol. 14, e1002476.

Arrowsmith, C.H., Audia, J.E., Austin, C., Baell, J., Bennett, J., 2015. The promise and peril of chemical probes. Nat. Chem. Biol. 11, 536–541.

Artus, M., van der Windt, D.A., Jordan, K.P., Hay, E.M., 2010. Low back pain symptoms show a similar pattern of improvement following a wide range of primary care treatments: a systematic review of randomized clinical trials. Rheumatology 49, 2346–2356.

Azoulay, P., Bonatti, A., Krieger, J.L., 2015. The Career Effects of Scandal: Evidence from Scientific Retractions. NBER Working Paper No. 21146. Available from: http://www.nber.org/papers/w21146.

Baggerly, K.A., Coombes, K.R., 2009. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. Annal. Appl. Stat. 3, 1309–1334.

Bahrndorff, S., Alemu, T., Alemneh, T., Nielsen, J.L., 2016. The microbiome of animals: implications for conservation biology. Int. J. Genomics 2016, 7.

Baker, M., 2015. Blame it on the antibodies. Nature 521, 274–276.

Baker, M., 2016a. Muddled meanings hamper efforts to fix reproducibility crisis. Nature. Available from: http://www.nature.com/news/muddled-meanings-hamper-efforts-to-fix-reproducibility-crisis-1.20076.

Baker, M., 2016b. Is there a reproducibility crisis? Nature 533, 452–454.

Baker, M., Dolgin, E., 2017. Reproducibility project yields muddy results. Nature 541, 269–270.

Balch, C., Arias-Pulido, H., Banerjee, S., Lancaster, A.K., Clark, K.B., Perilstein, M., et al., 2015. Science and technology consortia in U.S. biomedical research: a paradigm shift in response to unsustainable academic growth. Bioessays 37, 119–122.

Barber, M., 2014. Rab Butler's 1944 act brings free secondary education for all. BBC News School Report, January 17, 2014. Available from: http://www.bbc.co.uk/schoolreport/25751787.

Barrows, N.J., Le Sommer, C., Garcia-Blanco, M.A., Pearson, J.L., 2010. Factors affecting reproducibility between genome-scale siRNA-based screens. J. Biomol. Screening 15, 735–747.

Bartlett, T., 2015. Can a Longtime Fraud Help Fix Science? Chron. Higher Edu. Available from: http://chronicle.com/article/Can-a-Longtime-Fraud-Help-Fix/231061/.

Bath, P.M.W., Gray, L.J., Bath, A.J.G., Buchan, A., Miyata, T., Green, A.R., 2009. On behalf of the NXY-059 Efficacy Meta-analysis in individual Animals with Stroke (NEMAS) investigators. Effects of NXY-059 in experimental stroke: an individual animal meta-analysis. Br. J. Pharmacol. 157, 1157–1171.

Beall, J., 2012. Predatory publishers are corrupting open access. Nature 489, 179.

Beall, J., 2016. Beall's List of Predatory Publishers 2016. Scholary Open Access. Available from: http://scholarlyoa.com/2016/01/05/bealls-list-of-predatory-publishers-2016/.

Beall, J., 2017. What I learned from predatory publishers. Biochem. Med. 27, 273–278.

Begley, C.G., 2013. Reproducibility: six red flags for suspect work. Nature 497, 433–434.

Begley, C.G., 2017a. Quoted in Harris R. 2017. Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions. Basic Books, New York, p. 26.

Begley, C.G., 2017b. Quoted in Harris R. 2017. Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions. Basic Books, New York, p. 9.

Begley, C.G., Ellis, L.M., 2012. Drug development. Raise standards for preclinical cancer research. Nature 483, 531–533.

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in science: improving the standard for basic and preclinical research. Cir. Res. 116, 116–126.

Begley, S., 2012. In cancer science, many "discoveries" don't hold up. Reuters. Available from: http://www.reuters.com/article/2012/03/28/us-science-cancer-idUSBRE82R12P20120328.

Bell, C.J., Dinwiddie, D.L., Miller, N.A., Hateley, S.L., Ganusova, E.E., et al., 2011. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci. Transl. Med. 3, 65ra4.

Bhattacharjee, Y., 2013. The Mind of a Con Man, New York Times Magazine. Available from: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all&_r=1&.

Bilton, N., 2016a. The secret culprit in the Theranos mess. Vanity Fair. Available from: http://www.vanityfair.com/news/2016/05/theranos-silicon-valley-media.

Bilton, N., 2016b. Exclusive: How Elizabeth Holmes's House of Cards came tumbling down. Vanity Fair. Available from: http://www.vanityfair.com/news/2016/09/elizabeth-holmes-theranos-exclusive.

Bishop, D., 2012. Novelty, interest and replicability. Bishop Blog. Available from: http://deevybee.blogspot.co.uk/2012/01/novelty-interest-and-replicability.html.

Bishop, D., 2015. Publishing replication failures: some lessons from history. Bishop Blog. Available from: http://deevybee.blogspot.co.uk/2015/07/publishing-replication-failures-some.html.

Bishop, W.H., 2013. The role of ethics in 21st century organizations. J. Bus. Ethics 118, 635–637.

Blatt, M.R., 2015. Vigilante science. Plant Physiol. 169, 907–909.

Boettiger, S., Bennett, A.B., 2006. Bayh-Dole: if we knew then what we know now. Nat. Biotechnol. 24, 320–323.

Bohannon, J., 2013. Who's afraid of peer review? Science 342, 60–65.

Bohannon, J., 2015. How to hijack a journal. Science 350, 903–905.

Bollen, K., Cacioppo, J.T., Kaplan, R., Krosnick, J., Old, J.L., 2015. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science, National Science Foundation, Arlington, VA.

Bornemann-Cimenti, H., Szilagyi, I.S., Sandner-Kiesling, A., 2015. Perpetuation of retracted publications using the example of the Scott S. Reuben case: incidences, reasons and possible improvements. Sci. Eng. Ethics 22, 1063–1072.

Borrell, B., 2009. A medical madoff: anesthesiologist faked data in 21 studies. Sci. Amer. Available from: http://www.scientificamerican.com/article/a-medical-madoff-anesthestesiologist-faked-data/.

Borsboom, D., Wagenmakers, E.-J., 2013. Book Review. Derailed: The Rise and Fall of Diederik Stapel. APS Observer. Available from: http://www.psychologicalscience.org/index.php/publications/observer/2013/january-13/derailed-the-rise-and-fall-of-diederik-stapel.html.

Boston, M., 2015. Retracted Scientific Studies: A Growing List. New York Times. Available from: http://www.nytimes.com/interactive/2015/05/28/science/retractions-scientific.

Bothwell, L.E., Greene, J.A., Podolsky, S.H., Jones, D.S., 2016. Assessing the gold standard—lessons from the history of RCTs. N. Engl. J. Med. 374, 2175–2181.

Bradbury, A., Pluckthun, A., 2015. Standardize antibodies used in research. Nature 518, 27–29.

Braude, S.E., 1979. ESP and Psychokinesis: A Philosophical Examination. Temple University Press, Philadelphia, PA, p. 2.

Broad, W., Wade, N., 1983. Betrayers of the Truth. Touchstone/Simon and Shuster, New York, p. 12.

Buntin, M.B., Burke, M.F., Hoaglin, M.C., Blumenthal, D., 2011. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff. 30, 464–471.

Budd, J.M., Sievert, M., Schultz, T.R., Scoville, C., 1999. Effects of article retraction on citation and practice in medicine. Bull. Med. Libr. Assoc. 87, 437–443.

Bush, V., 1945. Science, the Endless Frontier: A Report to the President. US Government Printing Office, Washington, DC.

Butler, D., 2013. Investigating journals: the dark side of publishing. Nature 495, 433–435.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.

California Biomedical Research Association (CBRA), 2017. Fact Sheet What is Biomedical Research? Available from: http://www.ca-biomed.org/pdf/media-kit/fact-sheets/FS-WhatBiomedical.pdf.

Callahan, D., 2004. The Cheating Culture: Why More Americans Are Doing Wrong to Get Ahead. Harcourt, Orlando, FL.

Caldwell, G.W., Masucci, J.A., Yan, Z., Hageman, W., 2004. Allometric scaling of pharmacokinetic parameters in drug discovery: can human CL, Vss and $t_{1/2}$ be predicted from in vivo rat data? Eur. J. Drug Metabol. Pharmacokinet. 29, 133–143.

Carlisle, J.B., 2012. The analysis of 169 randomised controlled trials to test data integrity. Anaesthesia 67, 521–537.

Carey, B., 2015. Science, Now Under Scrutiny Itself. New York Times. Available from: http://www.nytimes.com/2015/06/16/science/retractions-coming-out-from-under-science-rug.html?_r=0\.

Carreyrou, J., 2015. Hot startup theranos has struggled with its blood-test technology. Wall St. J. Available from: http://www.wsj.com/articles/theranos-has-struggled-with-blood-tests-1444881901.

Carreyrou, J., 2016. Theranos whistleblower shook the company—and his family. Wall St. J. Available from: http://www.wsj.com/articles/theranos-has-struggled-with-blood-tests-1444881901.
Casadevall, A., Fang, F.C., 2010. Reproducible science. Infect. Immun. 78, 4792–4795.
Chalmers, I., Bracken, M.B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A.M., et al., 2014. How to increase value and reduce waste when research priorities are set. Lancet 383, 156–165.
Chau, C.H., Rixe, O., McLeod, H., Figg, W.D., 2008. Validation of analytical methods for biomarkers employed in drug development. Clin. Cancer Res. 14, 5967–5976.
Chawla, D.S., 2017. Mystery as controversial list of predatory publishers disappears. ScienceInsider. Available from: http://www.sciencemag.org/news/2017/01/mystery-controversial-list-predatory-publishers-disappears.
Chen, J.J., Hsueh, H.M., Delongchamp, R.R., Lin, C.J., Tsai, C.A., 2007. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics 8, 412.
Cheng, B.Q., Jia, C.Q., Liu, C.T., Fan, W., Wang, Q.L., Zhang, Z.L., Yi, C.H., 2008. Chemoembolization combined with radiofrequency ablation for patients with hepatocellular carcinoma larger than 3 cm: a randomized controlled trial. JAMA 299, 1669–1677.
Clooney, A.G., Fouhy, F., Sleator, R.D., O'Driscoll, A., Stanton, C., Cotter, P.D., Claesson, M.J., 2016. Comparing apples and oranges? Next generation sequencing and its impact on microbiome analysis. PLoS One 11, e0148028.
Colquhoun, D., 2011. Publish-or-perish: peer review and the corruption of science. Guardian, Available from: http://www.theguardian.com/science/2011/sep/05/publish-perish-peer-review-science.
Colquhoun, D., 2014. An investigation of the false discovery rate and the misinterpretation of p-values. R. Soc. Open. Sci. 1, 140216.
Collins, F.S., Tabak, L.A., 2014. Policy: NIH plans to enhance reproducibility. Nature 505, 612–613.
Comery, T.A., Martone, R.L., Aschmies, S., Atchison, K.P., Diamantidis, G., et al., 2005. Acute γ-secretase inhibition improves contextual fear conditioning in the Tg2576 mouse model of Alzheimer's disease. J. Neurosci. 25, 8898–8902.
Cook, D.J., Teves, L., Tymianski, M., 2012. A translational paradigm for the preclinical evaluation of the stroke neuroprotectant Tat-NR2B9c in Gyrencephalic nonhuman primates. Sci. Transl. Med. 4, 154ra133.
Coombes, K.R., Wang, Baggerly, K.A., 2007. Microarrays: retracing steps. Nat. Med. 13, 1226–1227.
Couzin-Frankel, J., 2017. Firing of veteran NIH scientist prompts protests over publication ban. Science. Available from: http://www.sciencemag.org/news/2017/02/firing-veteran-nih-scientist-prompts-protests-over-publication-ban.
Crabbe, J.C., Wahlsten, D., Dudek, B.C., 1999. Genetics of mouse behavior: interactions with lab environment. Science 284, 1670–1672.
Crocker, J., 2011. The road to fraud starts with a single step. Nature 479, 151.
Cryan, J.F., O'Mahoney, S.M., 2011. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterol. Motility 23, 187–192.
Curtis, M.J., Bond, R.A., Spina, D., Ahluwalia, A., Alexander, S.P.A., et al., 2015. Experimental design and analysis and their reporting: new guidance for publication in BJP. Br. J. Pharmacol. 172, 3461–3471.
Cyranoski, D., 2004. Korea's stem-cell stars dogged by suspicion of ethical breach. Nature 429, 3.
Cyranoski, D., 2014a. Cloning comeback. Nature 505, 468–471.
Cyranoski, D., 2014b. Stem-cell pioneer blamed media 'bashing' in suicide note. Nat. News. Available from: http://www.nature.com/news/stem-cell-pioneer-blamed-media-bashing-in-suicide-note-1.15715.
Cyranoski, D., Check, E., 2005. Clone star admits lies over eggs. Nature 438, 536–537.
Davis, P., 2011. Quoted in Mandavilli A. Peer review: trial by twitter. Nature 469, 286–287.
Decullier, E., Huot, L., Samson, G., Maisonneuve, H., 2013. Visibility of retractions: a cross-sectional one-year study. BMC Res. Notes 6, 238.
Decullier, E., Huot, L., Maisonneuve, H., 2014. What time lag for a retraction search on PubMed? BMC Res. Notes 7, 395.
De Los Angeles, A., Ferrari, F., Fujiwara, Y., Mathieu, R., Lee, S., Lee, S., et al., 2015. Failure to replicate the STAP cell phenomenon. Nature 525, E6–E9.
Deer, B., 2010. Wakefield's "autistic enterocolitis" under the microscope. BMJ 340, c1127.
Deer, B., 2011a. How the case against the MMR vaccine was fixed. BMJ 342, c5347.
Deer, B., 2011b. How the vaccine crisis was meant to make money. BMJ 342, c5258.
Denayer, T., Stöhr, T., Van Roy, M., 2014. Animal models in translational medicine: validation and prediction. New Horiz. Transl. Med. 2, 5–11.

Diener, H.C., Lees, K.R., Lyden, P., Grotta, J., Davalos, A., Davis, S.M., et al., 2008. NXY-059 for the treatment of acute stroke: pooled analysis of the SAINT I and II Trials. Stroke 39, 1751–1758.

Dinan, T.G., Roman, M., Stilling, R.M., Stanton, C., Cryan, J.F., 2015. Collective unconscious: how gut microbes shape human behavior. J. Psychiat. Res. 63, 1–9.

Dirnagl, U., Lauritzeb, M., 2010. Fighting publication bias: introducing the negative results section. J. Cereb. Blood Flow Metab. 30, 1263–1264.

Dirnagl, U., Macleod, M.R., 2009. Stroke research at a road block: the streets from adversity should be paved with meta-analysis and good laboratory practice. Br. J. Pharmacol. 157, 1154–1156.

Dolgin, E., 2014. Drug discoverers chart path to tackling data irreproducibility. Nat. Rev. Drug Discov. 13, 875–876.

Doyle, A., McGarry, M.P., Lee, N.A., Lee, J.J., 2012. The construction of transgenic and gene knockout/knockin mouse models of human disease. Transgenic Res. 21, 327–349.

Drummond, C., 2009. Replicability is not reproducibility: nor is it good science. Available from: http://www.site.uottawa.ca/ICML09WS/papers/w2.pdf.

Drucker, D.J., 2016. Never waste a good crisis: confronting reproducibility in translational research. Cell Metab 24, 348–360.

Dzeng, E., 2014. How academia and publishing are destroying scientific innovation: a conversation with Sydney Brenner. King's Rev. Available from: http://kingsreview.co.uk/magazine/blog/2014/02/24/how-academia-and-publishing-are-destroying-scientific-innovation-a-conversation-with-sydney-brenner/.

Earp, B.D., Trafimow, D., 2016. Replication, falsification, and the crisis of confidence in social psychology. Front. Psychol. 6, 621, 2015.

Economist, 2013a. Unreliable research. Trouble at the lab. The Economist. Available from: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Economist, 2013b. Looks good on paper. Available from: http://www.economist.com/news/china/21586845-flawed-system-judging-research-leading-academic-fraud-looks-good-paper.

Editors of the Lancet, 2010. Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 375, 445.

Engber, D., 2016. Cancer research is broken. Slate. Available from: http://www.slate.com/articles/health_and_science/future_tense/2016/04/biomedicine_facing_a_worse_replication_crisis_than_the_one_plaguing_psychology.html.

Errington, T.M., Iorns, E., Gunn, W., Tan, F.E., Lomax, J., Nosek, B.A., 2014. Science Forum: an open investigation of the reproducibility of cancer biology research. eLife 3, e04333.

Estes, S., 2012. The myth of self-correcting science. Atlantic Magazine. Available from: http://www.theatlantic.com/health/archive/2012/12/the-myth-of-self-correcting-science/266228/.

Ezenwa, V.O., Gerardo, N.M., Inouye, D.W., Medina, M., Xavier, J.B., 2012. Animal behavior and the microbiome. Science 338, 198–199.

Fairbanks, C.A., Goracke-Postle, C.J., 2015. Neurobiological studies of chronic pain and analgesia: rationale and refinements. Eur. J. Pharmacol. 759, 168–181.

Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. Natl. Sci. Rev. 1, 293–314.

Fanelli, D., 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One 4, e5738.

Fanelli, D., 2012. Negative results are disappearing from most disciplines and countries. Scientometrics 90, 891–904.

Fanelli, D., Ioannidis, J.P.A., 2013. US studies may overestimate effect sizes in softer research. Proc. Natl. Acad. Sci. USA 110, 15031–15036.

Fang, F.C., Casadevall, A., 2011. Retracted science and the retraction index. Infect. Immun. 79, 3855–3859.

Fang, F.C., Steen, R.G., Casadevall, A., 2012. Misconduct accounts for the majority of retracted scientific publications. Proc. Natl. Acad. Sci. USA 109, 17028–17033.

Farrell, M.S., Werge, T., Sklar, P., Owen, M.J., Ophoff, R.A., O'Donovan, M.C., et al., 2015. Evaluating historical candidate genes for schizophrenia. Mol. Psychiatr. 20, 555–562.

Faulkes, Z., 2016. Mission creep in scientific publishing. NeuroDojo. Available from: http://neurodojo.blogspot.com/2016/02/mission-creep-in-scientific-publishing.html?m=1.

Fishburn, C.S., 2014. Repairing reproducibility. SciBx 7. Available from: http://www.nature.com/scibx/journal/v7/n10/full/scibx.2014.275.html.

Fisher, K.H., Wright, V.M., Taylor, A., Zeidler, M.P., Brown, S., 2012. Advances in genome-wide RNAi cellular screens: a case study using the *Drosophila* JAK/STAT pathway. BMC Genomics 13, 506.

Flam, F., 2016. Lesson of theranos: fact-checking alone isn't enough. BloombergView. Available from: http://www.bloomberg.com/view/articles/2016-08-08/lesson-of-theranos-fact-checking-alone-isn-t-enough.
Flier, J.S., 2017. Irreproducibility of published bioscience research: diagnosis, pathogenesis and therapy. Mol. Metab. 6, 2–9.
Freedman, D.H., 2010. Lies, damned lies, and medical science. Atlantic Monthly. Available from: http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/.
Freedman, L.P., Gibson, M.C., 2015. The impact of preclinical irreproducibility on drug development. Clin. Pharmacol. Ther. 97, 16–18.
Freedman, L.P., Inglese, J., 2014. The increasing urgency for standards in basic biologic research. Cancer Res. 74, 4024–4029.
Freedman, L.P., Cockburn, I.M., Simcoe, T.S., 2015. The economics of reproducibility in preclinical research. PLoS Biol. 13, e1002165.
Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., et al., 2016. The genetic architecture of type 2 diabetes. Nature 536, 41–47.
Geraghty, R.J., Capes-Davis, A., Davis, J.M., Downward, J., Freshney, R.I., Knezevic, I., et al., 2014. Guidelines for the use of cell lines in biomedical research. Br. J. Cancer 111, 1021–1046.
Gewin, V., 2014. Retractions: a clean slate. Nature 507, 389–391.
Gibson, S., 2012. Don't tar discipline with Stapel brush. Times Higher Edu. Available from: https://www.timeshighereducation.com/dont-tar-discipline-with-stapel-brush/422194.article.
Gilbert, D.T., King, G., Pettigrew, S., Wilson, T.D., 2016. Comment on "estimating the reproducibility of psychological science". Science 351, 1037a.
Gillis, M., 2017. U.S. company launches a new blacklist of deceptive academic journals. University Affairs. Available from: http://www.universityaffairs.ca/news/news-article/u-s-company-launches-new-blacklist-deceptive-academic-journals/.
Glass, D.J., 2014. Experimental Design for Biologists, second ed. Cold Spring Harbor Press, New York, Cold Spring Harbor.
Global Lipids Genetics Consortium, 2013. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283.
Goodlee, F., Smith, J., Marcovitch, H., 2011. Wakefield's article linking MMR vaccine and autism was fraudulent. BMJ 342, c7452.
Goldacre, B., 2012. Bad Pharma. London, Fourth Estate, 2012.
Goldberg, P., 2015. Duke officials silenced med student who reported trouble in Anil Potti's Lab. Cancer Lett. Available from: http://www.cancerletter.com/articles/20150109_1.
Goodman, S.N., Fanelli, D., Ioannidis, J.P.A., 2016. What does research reproducibility mean? Sci. Transl. Med. 8, 342ps12.
Goodstein, D., 1995. Conduct and misconduct in science. Ann. NY Acad. Sci. 775, 31–38.
Goodyear, D., 2016. The stress test. Rivalries, intrigue, and fraud in the world of stem-cell research. New Yorker. Available from: http://www.newyorker.com/magazine/2016/02/29/the-stem-cell-scandal.
Gorski, D., 2011. The wrong way to "open up" clinical trials. Science-Based Medicine. Available from: https://www.sciencebasedmedicine.org/the-wrong-way-to-open-up-clinical-trials/.
Gorski, D., 2012. The problem with preclinical research? Or: a former pharma exec discovers the nature of science. Science-Based Medicine. Available from: https://www.sciencebasedmedicine.org/the-problem-with-preclinical-research/.
Gorski, D., 2016. Is there a reproducibility "crisis" in biomedical science? No, but there is a reproducibility problem. Science-Based Medicine. Available from: https://www.sciencebasedmedicine.org/is-there-a-reproducibility-crisis-in-biomedical-science-no-but-there-is-a-reproducibility-problem/.
Greek, R., Hansen, L.A., 2013. Questions regarding the predictive value of one evolved complex adaptive system for a second: exemplified by the SOD1 mouse. Prog. Biophys. Mol. Biol. 113, 231–253.
Groenink, L., Folkerts, G., Schuurman, H.-J., 2015. European Journal of Pharmacology, special issue on translational value of animal models: introduction. Eur. J. Pharmacol. 759, 1–2.
Gross, G., 2016. Scientific Misconduct. Annu.Rev. Psychol. 67, 693–711.
Grove, A.S., 2005. Efficiency in the health care industries: a view from the outside. JAMA 294, 490–492.
Grove, A., 2011. Rethinking clinical trials. Science 333, 1679.
Harris, R., 2017. Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions. Basic Books, New York.

Harrison, R.K., 2016. Phase II and phase III failures: 2013-2015. Nat. Rev. Drug Discov. 15, 817–818.
Hartung, T., 2013. Food for thought, look back in anger—What clinical studies tell us about preclinical work. ALTEX 30, 275–291.
Hatzipetros, T., Bogdanik, L.P., Tassinari, V.R., Kidd, J.D., Moreno, A.J., Davis, C., et al., 2014. C57BL/6J congenic Prp-TDP43A315T mice develop progressive neurodegeneration in the myenteric plexus of the colon without exhibiting key features of ALS. Brain Res. 1584, 59–72.
Heggeness, M.L., Gunsalus, K.T.W., Pacas, J.G., McDowell, G., 2017. The new face of US science. Nature 541, 21–23.
Henderson, V.C., Kimmelman, J., Fergusson, D., Grimshaw, J.M., Hackam, D.G., 2013. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 10, e1001489.
Hill, M.D., Martin, R.H., Mikulis, D., Wong, J.H., Silver, F.L., ter Brugge, K.G., et al., 2012. Safety and efficacy of NA-1 in patients with iatrogenic stroke after endovascular aneurysm repair (ENACT): a phase 2, randomised, double-blind, placebo-controlled trial. Lancet Neurol. 11, 942–950.
Hiltzik, M., 2014. More on the unsavory history of the vaccine-autism 'link'. LA Times, Available from: http://www.latimes.com/business/hiltzik/la-fi-mh-vaccineautism-link-20140122,0,1151028.story#axzz2rXPki7fV.
Hirst, J.A., Howick, J., Aronson, J.K., Roberts, N., Perera, R., Koshiaris, C., et al., 2014. The need for randomization in animal trials: an overview of systematic reviews. PLoS One 9, e98856.
Hobin, J.A., Deschamps, A.M., Bockman, R., Cohen, S., Dechow, P., Eng, C., et al., 2012. Engaging basic scientists in translational research: identifying opportunities, overcoming obstacles. J. Transl. Med. 10, 72.
Hodgson, J., 2016. When biotech goes bad. Nat. Biotech. 34, 284–291.
Hoffman, J., 2016. Vaxxed review—one-sided film leaves the elephant in the room. Guardian. Available from: https://www.theguardian.com/film/2016/apr/02/vaxxed-from-cover-up-to-catastrophe-review.
Holder, D.J., Marino, M.J., 2017. Enhancing reproducibility: logic in experimental design and execution in pharmacology and drug discovery. Curr. Protocol Pharmacol. 76, A.3G.1–A.3G.26.
Hongo, J., 2014. Timeline: the rise and fall of Haruko Obokata in 2014. Wall St. J., Japan JAPANREALTIME. Available from: http://blogs.wsj.com/japanrealtime/tag/haruko-obokata/.
Horrabin, D.F., 2001. Something rotten at the core of science? Trends Pharmacol. Sci. 22, 51–52.
Horrabin, D., 2003. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? Nat. Rev. Drug Discov. 2, 151–154.
Horton, R., 2000. Genetically modified food: consternation, confusion, and crack-up. Med. J. Aust. 172, 148–149.
Horton, R., 2015. Offline: what is medicine's 5 sigma? Lancet 285, 1380.
Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., et al., 2008. The future of biocuration. Nature 455, 47–50.
Hróbjartsson, A., Gøtzsche, P.C., 2010. Placebo interventions for all clinical conditions. Cochrane Database Syst. Rev. 20 (1), CD003974.
Hsu, J.C., Chang, J., Wang, T., Steingrímsson, E., Magnússon, M.K., Bergsteinsdottir, K., 2007. Statistically designing microarrays and microarray experiments to enhance sensitivity and specificity. Brief Bioinform. 8, 22–31.
Hwang, W.S., Ryu, Y.J., Park, J.H., Park, E.S., Lee, E.G., Koo, J.M., et al., 2004. Evidence of a pluripotent human embryonic stem cell line derived from a cloned blastocyst. Science 303, 1669–1674.
Hwang, W.S., Roh, S.I., Lee, B.C., Kang, S.K., Kwon, D.K., Kim, S., et al., 2005. Patient-specific embryonic stem cells derived from human SCNT blastocysts. Science 306, 1777–1783.
Interlandi, J., 2006; An Unwelcome Discovery. New York Times. Available from: http://www.nytimes.com/2006/10/22/magazine/22sciencefraud.html.
Ioannidis, J.P.A., 2005. Why most published research findings are false. PLoS Med. 2, e124.
Ioannidis, J.P.A., 2006. Evolution and translation of research findings: from bench to where? PLoS Clin. Trials 1, e36.
Ioannidis, J.P.A., 2014. How to make more published research true. PLoS Med. 11, e1001747.
Ioannidis, J.P.A., 2015. Stealth research. Is biomedical innovation happening outside the peer-reviewed literature? JAMA 313, 663–664.
Ioannidis, J.P.A., 2016a. Why most clinical research is not useful. PLoS Med. 13, e1002049.
Ioannidis, J.P.A., 2016b. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Quart. 94, 485–545.
Ioannidis, J.P.A., Khoury, M.J., 2014. Assessing value in biomedical research: the PQRST of appraisal and reward. JAMA 312, 483–484.

Ioannidis, J.P., Trikalinos, T.A., 2005. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. J. Clin. Epidemiol. 58, 543–549.
Ioannidis, J.P.A., Allison, D.B., Ball, C.A., Coulibaly, I., Cui, X., et al., 2009. Repeatability of published microarray gene expression analyses. Nat. Genetics 41, 149–155.
Iizuka, M., Chiba, M., Yukawa, M., Nakagomi, T., Fukushima, T., Watanabe, S., Nakagomi, O., 2000. Immunohistochemical analysis of the distribution of measles related antigen in the intestinal mucosa in inflammatory bowel disease. Gut 46, 163–169.
Jain, A., Marshall, J., Buikema, A., Bancroft, T., Kelly, J.P., Newschaffer, C.J., 2015. Autism occurrence by MMR vaccine status among US children with older siblings with and without Autism. JAMA 313, 1534–1540, 2015.
Jarvis, M.F., Williams, M., 2016. Irreproducibility in preclinical biomedical research: perceptions, uncertainties and knowledge gaps. Trends Pharmacol. Sci. 37, 290–302.
Jian, C., Zou, D., Liu, X., Meng, L., Huang, J., et al., 2016. Cognitive deficits are ameliorated by reduction in amyloidβ accumulation in Tg2576/p75(NTR+/−) mice. Life Sci. 155, 167–173.
Jiji, 2016. Obokata sticks to guns, launches website with instructions for making STAP cells. Japan Times. Available from: http://www.japantimes.co.jp/news/2016/04/01/national/science-health/obokata-sticks-guns-launches-website-boasting-way-make-stap-cells/#.V3MzeVfKKeN.
Jobe, P.C., Adams-Curtis, L.E., Burks, T.F., Fuller, R.W., Peck, C.C., Ruffolo, R.R., Snead, O.C., et al., 1994. The essential role of integrative bio-medical sciences in protecting and contributing to the health and well-being of our nation. Physiologist 37, 79–86.
Jump, P., 2011. A star's collapse. Times Higher Education. Available from: https://www.insidehighered.com/news/2011/11/28/scholars-analyze-case-massive-research-fraud.
Kaiser, M., 2014. The integrity of science—lost in translation? Best Pract. Res. Clin. Gastroenterol. 28, 339–347.
Kahn, R.S., Sommer, I.E., Murray, R.M., Meyer-Lindenberg, A., Weinberger, D.R., Cannon, T.D., et al., 2015. Schizophrenia. Nat. Rev. Dis. Primers 1, 15067.
Kakluk, P., 2009. The legacy of the Hwang case: research misconduct in biosciences. Sci. Eng. Ethics 15, 545–562.
Kass, R.E., Caffo, B.S., Davidian, M., Meng, X.-L., Yu, B., Reid, N., 2016. Ten simple rules for effective statistical practice. PLoS Comput. Biol. 12, e1004961.
Kenakin, T., Bylund, D.B., Toews, M.L., Mullane, K., Winquist, R.J., Williams, M., 2014. Replicated, replicable and relevant—target engagement and pharmacological experimentation in the 21st Century. Biochem. Pharmacol. 87, 64–77.
Kerr, N.L., 1998. HARKing: hypothesizing after the results are known. Pers. Soc. Psychol. Rev. 2, 196–217.
Kimmelman, J., Federico, C., 2017. Consider drug efficacy before first-in-human trials. Nature 542, 25–27.
Kleiman, R.J., Ehlers, M.D., 2016. Data gaps limit the translational potential of preclinical research. Sci. Transl. Med. 8, 320 ps1.
Kochan, C.A., Budd, J.M., 1992. The persistence of fraud in the literature: the Darsee case. J. Am. Soc. Info. Sci. 43, 488–493.
Kolata, G., 2011. How bright promise in cancer testing fell apart. NY Times. Available from: http://www.nytimes.com/2011/07/08/health/research/08genes.html?_r=0.
Kornfeld, D.S., 2012. Perspective: research misconduct: the search for a remedy. Acad. Med. 87, 877–882.
Kornfeld, D.S., Titus, S.L., 2016. Stop ignoring misconduct. Nature 537, 29–30.
Kossoff, M., 2017. Theranos's latest lawsuit may be its worst yet. Vanity Fair. Available from: http://www.vanityfair.com/news/2017/01/theranoss-latest-lawsuit-may-be-its-worst-yet.
Kotecki, P., 2016. In focus: as Lyrica profits dry up, Northwestern seeks another 'blockbuster' drug. dailynorthwestern.com. Available from: http://dailynorthwestern.com/2016/04/10/in-focus/in-focus-as-lyrica-profits-dry-up-northwestern-seeks-another-blockbuster-drug/.
Kranke, P., Apfel, C.C., Roewer, N., 2000. Reported data on Granisetron and postoperative nausea and vomiting by Fujii et al. are incredibly nice! Anesth. Anal. 90, 1004–1006.
Kraus, W.L., 2014. Editorial: do you see what i see? Quality, reliability, and reproducibility in biomedical research. Mol. Endocrinol. 38, 277–280.
Lau, J., Ioannidis, J.P.A., Terrin, N., Schmid, C.H., Olkin, I., 2006. The case of the misleading funnel plot. BMJ 333, 597–600.
Lee, B.C., Kim, M.K., Jang, G., Oh, H.J., Yuda, F., Kim, H.J., et al., 2005. Dogs cloned from adult somatic cells. Nature 436, 641.
Lehrer, J., 2010. Annals of science. The truth wears off. Is There something wrong with the scientific method? New Yorker. Available from: http://archives.newyorker.com/?i=2010-12-13#folio=052.

Lemaitre, B., 2016. An essay on science and narcissism: how do high-ego personalities drive research in life sciences? brunolemaitre.ch, Switzerland, Lausanne. Available from: https://www.amazon.com/Essay-Science-Narcissism-high-ego-personalities-ebook/dp/B01DS47AN4.

Levelt Committee, 2011. Interim report regarding the breach of scientific integrity by Prof. D. A. Stapel. Tilburg University. Available from: https://www.tilburguniversity.edu/upload/547aa461-6cd1-48cd-801b-61c434a73f79_interim-report.pdf.

Llovera, G., Hofmann, K., Roth, S., Salas-Pérdomo, A., Ferrer-Ferrer, M., Perego, C., et al., 2015. Results of a preclinical randomized controlled multicenter trial (pRCT): anti-CD49d treatment for acute brain ischemia. Sci. Transl. Med. 7, 299ra121.

Longo, D.L., Drazen, J.M., 2016. Data sharing. N. Engl. J. Med. 374, 276–277.

Loscalzo, J., 2012. Irreproducible experimental results: causes, (mis) interpretations, and consequences. Circulation 125, 1211–1214.

Lowe, D., 2011. Andy Grove's idea for opening up clinical trials. In the pipeline. Sci. Transl. Med. Available from: http://blogs.sciencemag.org/pipeline/archives/2011/09/28/andy_groves_idea_for_opening_up_clinical_trials.

Lu, S.F., Jin, G.Z., Uzzi, B., Jones, B., 2013. The retraction penalty: evidence from the web of science. Sci. Rep. 3, 3146.

Lucanic, M., Plummer, W.T., Chen, E., Harke, J., Foulger, A.C., Onken, B., et al., 2017. Impact of genetic background and experimental reproducibility on identifying chemical compounds with robust longevity effects. Nat. Commun. 8, 14256.

Lukk, M., Kapushesky, M., Nikkilä, J., Parkinson, H., Goncalves, A., et al., 2010. A global map of human gene expression. Nat. Biotech. 28, 322–324.

Macleod, M.R., Van der Worp, B., Sena, E.S., Howells, D.W., Dirnagl, U., Donnan, G.A., 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 39, 2824–2829.

Macleod, M.R., Fisher, M., O'Collins, V., Sena, E.S., Dirnagl, U., Bath, P.M., et al., 2009. Good laboratory practice: preventing introduction of bias at the bench. J. Cereb. Blood Flow Metab. 29, 221–223.

Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J.P.A., et al., 2014. Biomedical research: increasing value, reducing waste. Lancet 383, 101–104, 2014.

Macrae, I.M., 2011. Preclinical stroke research—advantages and disadvantages of the most common rodent models of focal ischaemia. Br. J. Pharmacol. 164, 1062–1078.

Majumder, M.S., Cohn, E.L., Mekaru, S.R., Huston, J.E., Brownstein, J.S., 2015. Substandard vaccination compliance and the 2015 measles outbreak. JAMA Pediatr. 169, 494–495.

Marcus, A., Oransky, I., 2011. Science publishing: the paper is not sacred. Nature 480, 449–450.

Marcus, A., Oransky, I., 2014. What studies of retractions tell us. J. Microbiol. Biol. Edu. 15, 151–154.

Marcus, A., Oransky, I., 2015. How the biggest fabricator in science got caught. Nautilus. Available from: http://nautil.us/issue/24/error/how-the-biggest-fabricator-in-science-got-caught.

Marcus, A., McCook, A., Oransky, I., 2015. The top 10 retractions of 2015. The Scientist. Available from: http://www.the-scientist.com/?articles.view/articleNo/44895/title/The-Top-10-Retractions-of-2015/.

Marino, M., 2014. The use and misuse of statistical methodologies in pharmacology research. Biochem. Pharmacol. 87, 78–92.

Markel, H., 2013. Patents, Profits, and the American People—The Bayh–Dole Act of 1980. N. Engl. J. Med. 369, 794–796.

Mathur, M., 2016. Replication of "Why People are Reluctant to Tempt Fate" by Risen & Gilovich (2008, J. Personal. Social Psychol.) Risen & Gilovich replication writeup.pdf (Version: 1). OSC. Available from: https://osf.io/nwua6/.

Matosin, N., Frank, E., Engel, M., Lum, J.S., Newell, K.A., 2014. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. Dis. Model. Mech. 7, 171–173.

McClain, S., 2013. Not breaking news: many scientific studies are ultimately proved wrong! Guardian. Available from: http://www.theguardian.com/science/occams-corner/2013/sep/17/scientific-studies-wrong.

McGonigle, P., Ruggeri, B., 2014. Animal models of human disease: challenges in enabling translation. Biochem. Pharmacol. 87, 162–171.

McGonigle, P., Williams, M., 2014. Preclinical pharmacology and toxicology - contributions to the translational interface. Ref Module Biomed Sci. Available from: http://dx.doi.org/10.1016/B978-0-12-801238-3.05242-9.

McGorry, R.W., Webster, B.S., Snook, S.H., Hsiang, S.M., 2000. The relation between pain intensity, disability, and the episodic nature of chronic and recurrent low back pain. Spine 25, 834–841.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., et al., 2015. The human transcriptome across tissues and individuals. Science 348, 660–665.

Menezes, P., Miller, W.C., Wohl, D.A., Adimora, A.A., Leone, P.A., Miller, W.C., Eron, Jr., J.J., 2011. Does HAART efficacy translate to effectiveness? Evidence for a trial effect. PLos One 6 (7), e21824.

Michael, A., 2015. Ask The Chefs: How Can We Improve the Article Review and Submission Process? the scholarly kitchen. Available from: http://scholarlykitchen.sspnet.org/2015/03/26/ask-the-chefs-how-can-we-improve-the-article-review-and-submission-process/.

Mirowski, P., 2012. The Modern Commercialization of Science is a Passel of Ponzi Schemes. Social Epistemol. 26, 285–310.

Mobley, A., Linder, S.K., Braeuer, R., Ellis, L.M., Zwelling, L., 2013. A survey of data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. PloS One 8, e6322.

Moher, D., Glasziou, P., Chalmers, I., Nasser, M., Bossuyt, P.M.M., Korevaar, D.A., et al., 2016. Increasing value and reducing waste in biomedical research: who's listening? Lancet 397, 1573–1586.

Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segrè, A.V., et al., 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 981–990.

Morrison, S.J., 2014. Reproducibility project: cancer biology: time to do something about reproducibility. eLife 3, e03981.

Motulsky, H.J., 2014. Common misconceptions about data analysis and statistics. J. Pharmacol. Exp. Ther. 351, 200–205.

Moyé, L.A., Deswal, A., 2002. The fragility of cardiovascular clinical trial results. J. Card. Fail. 8, 247–253.

Muhlhausler, B.S., Bloomfield, F.H., Gillman, M.W., 2013. Whole animal experiments should be more like human randomized controlled trials. PLoS Biol. 11, e1001481.

Mullane, K., Williams, M., 2015. Unknown unknowns in biomedical research: does an inability to deal with ambiguity contribute to issues of irreproducibility? Biochem. Pharmacol. 97, 133–136.

Mullane, K., Williams, M., 2017. Enhancing reproducibility: failures from reproducibility initiatives underline core challenges. Biochem. Pharmacol. 138, 7–18.

Mullane, K., Enna, S.J., Piette, J., Williams, M., 2015. Guidelines for manuscript submission in the peer-reviewed pharmacological literature. Biochem. Pharmacol. 97, 224–239.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., et al., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1 Article no. 0021.

Nardone, R.M., 2008. Curbing rampant cross-contamination and misidentification of cell lines. BioTechniques 45, 221–227.

National Academies Press (NAP), 2015. Institute for Laboratory Animal Research. Round Table on Science and Welfare in Laboratory Animal Use. Reproducibility Issues in Research with Animals and Animal Models: Workshop in Brief. Washington, DC, The National Academies Press. Available from: http://www.nap.edu/catalog/21835/reproducibility-issues-in-research-with-animals-and-animal-models-workshop.

National Academies of Science (NASEM), 2017. Engineering and Medicine. Fostering Integrity in Research. Washington, DC, The National Academies Press. Available from: http://www.nap.edu/21896.

Nature, 2012. Editorial note. Nature 485, 41.

Nature, 2013. Time to talk. Online discussion is an essential aspect of the post-publication review of findings. Nature 502, 593–594.

Nature Editorial, 2016. Repetitive flaws. Nature 529, 256.

Nature Medicine Editorial, 2016. Take the long view. Nat. Med. 22, 1.

Neimark, J., 2014. The dirty little secret of cancer research. Discover. Available from: http://discovermagazine.com/2014/nov/20-trial-and-error.

Neimark, J., 2015. Line of attack. Science 347, 938–940.

Ni, J., Koyuturk, M., Tong, H., Haines, J., Xu, R., Zhang, X., 2016. Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. BMC Informatics 17, 453.

Nisbet, M., Markowitz, E.M., 2014. Understanding public opinion in debates over biomedical research: looking beyond political partisanship to focus on beliefs about science and society. PLoS One 9, e88473.

Normile, D., Vogel, G., 2014. STAP cells succumb to pressure. Science 344, 1215–1216.

Nosek, B.A., Errington, T.M., 2017. Making sense of replications. eLife 6, e23383.

Nutt, A.E., 2016. The scientist nearly went to jail for making up data. Washington Post. Available from: https://www.washingtonpost.com/news/speaking-of-science/wp/2016/04/01/when-scientists-lie-about-their-research-should-they-go-to-jail/?utm_term=.6bf002c03709.

O'Bryant, S.E., Waring, S.C., Cullum, C.M., Hall, J., Lacritz, L., et al., 2008. Staging dementia using clinical dementia rating scale sum of boxes scores: a Texas Alzheimer's research consortium study. Arch. Neurol. 65, 1091–1095.

O'Collins, V.E., Macleod, M.R., Donnan, G.A., Horky, L.L., van der Worp, B.H., Howells, D.W., 2006. 1,026 experimental treatments in acute stroke. Ann. Neurol. 59, 467–477.

Obokata, H., Wakayama, T., Sasai, Y., Kojima, K., Vacanti, M.P., Niwa, H., et al., 2014a. Retracted: stimulus-triggered fate conversion of somatic cells into pluripotency. Nature 505, 641–647.

Obokata, H., Sasai, Y., Niwa, H., Kadota, M., Andrabi, M., Takata, N., et al., 2014b. Retracted: bidirectional developmental potential in reprogrammed cells with acquired pluripotency. Nature 505, 676–680.

Omenn, G.S., Lane, L., Lundberg, E.K., Beavis, R.C., Nesvizhskii, A.I., Deutsch, E.W., 2015. Metrics for the human proteome project 2015: progress on the human proteome and guidelines for high-confidence protein identification. J. Proteome Res. 14, 3452–3460.

OSC (Open Science Collaboration), 2015. Estimating the reproducibility of psychological science. Science 349, aac4716.

Panagiotou, O.A., Willer, C.J., Hirschhorn, J.N., Ioannidis, J.P.A., 2013. The power of meta-analysis in genome-wide association studies. Annu. Rev. Genomics Hum. Genet. 14, 441–465.

Pattinson, D., 2012. Plos One launches reproducibility initiative. Available from: http://blogs.plos.org/everyone/2012/08/14/plos-one-launches-reproducibility-initiative/.

Peres-Neto, P.R., 2016. Will technology trample peer review in ecology? Ongoing issues and potential solutions. Okios 125, 3–9.

Perrin, S., 2014. Preclinical research: make mouse studies work. Nature 507, 423–425.

Peters, S.M., Pothuizen, H.H.J., Spruijt, B.M., 2015. Ethological concepts enhance the translational value of animal models. Eur. J. Pharmacol. 759, 42–50.

Piwowar, H.A., Vision, T.J., Whitlock, M.C., 2011. Data archiving is a good investment. Nature 473, 285.

Pfeifer, M.P., Snodgrass, G.L., 1990. The continued use of retracted, invalid scientific literature. JAMA 263, 1420–1423.

Pfeiffer, T., Bertram, L., Ioannidis, J.P.A., 2011. Quantifying selective reporting and the proteus phenomenon for multiple datasets with similar bias. PLoS One 6, e18362.

Potti, A., Dressman, H.K., Bild, A., Riedel, R.F., Chan, G., Sayer, R., et al., 2006. Retracted: genomic signatures to guide the use of chemotherapeutics. Nat. Med. 12, 1294–1300.

Price, A.L., Spencer, C.C.A., Donnelly, P., 2015. Progress and promise in understanding the genetic basis of common diseases. Proc. R. Soc. B 282, 20151684.

Prinz, F., Schlange, T., Asadullah, K., 2011. Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10, 712–713.

Prud'Homme, A., 2004. The cell game. Sam Waksal's Fast Money and False Promises—and the Fate of Imclone's Cancer Drug. Harper Business, New York.

Pulverer, B., quoted in Meadows A. "Research Mechanics", OA, Ethics, and More: Three Chefs Musings on APE 2016. The Scholarly Kitchen. Available from: https://scholarlykitchen.sspnet.org/2016/02/03/research-mechanics-oa-ethics-and-more-three-chefs-musings-on-ape-2016/.

Pusztai, L., Hatzis, C., Andre, F., 2013. Reproducibility of research and preclinical validation: problems and solutions. Nat. Rev. Clin. Oncol. 10, 720–724.

Ramkumar, A., 2015. Duke lawsuit involving cancer patients linked to Anil Potti settled. Duke Chronicle. Available from: http://www.dukechronicle.com/article/2015/05/duke-lawsuit-involving-cancer-patients-linked-anil-potti-settled.

Rasko, J., Power, C., 2015. What pushes scientists to lie? The disturbing but familiar story of Haruko Obokata. Available from: http://www.theguardian.com/science/2015/feb/18/haruko-obokata-stap-cells-controversy-scientists-lie.

Redman, B.K., 2015. Are the Biomedical Sciences Sliding Toward Institutional Corruption? And Why Didn't We Notice It? Edmond, J. Safra Working Papers, No. 59. Harvard University, 2015. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2585141.

Rennie, D., 1986. Guarding the guardians A conference on editorial peer review. JAMA 256, 2391–2392.

Rennie, D., Flanagin, A., Smith, R., Smith, J., 2003. Fifth international congress on peer review and biomedical publication. Call for research. JAMA 289, 1438.

Retraction Watch, 2016. Lawsuit against Ole Miss for rescinded Sarkar job offer dismissed; briefs filed in PubPeer case. RetractionWatch. Available from: http://retractionwatch.com/2016/01/20/judge-dismissed-lawsuit-against-ole-miss-for-rescinded-offer/.

Rogers, G.B., Kozlowska, J., Keeble, J., Metcalfe, K., Fao, M., Dowd, S.E., et al., 2014. Functional divergence in gastrointestinal microbiota in physically-separated genetically identical mice. Sci. Rep. 4, 5437.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638–641.
Roth, K.A., Cox, A.E., 2015. Science isn't science if it isn't reproducible. Am. J. Pathol. 185, 2–3.
Rung, J., Brazma, A., 2013. Reuse of public genome-wide gene expression data. Nat. Rev. Genet. 14, 89–99.
Rupp, B., Wlodawer, A., Minor, W., Helliwell, J.R., Jaskolski, M., 2016. Correcting the record of structural publications requires joint effort of the community and journal editors. FEBS J. 283, 4452–4457.
Ruxton, G.D., Colegrave, N., 2011. Experimental Design in the Life Sciences. Oxford University Press, Oxford, UK.
Salmon, D.A., McIntyre, C.R., Omer, S.B., 2015. Making mandatory vaccination truly compulsory: well intentioned but ill conceived. Lancet Infect. Dis. 15, 872–873.
Sams-Dodd, F., 2006. Strategies to optimize the validity of disease models in the drug discovery process. Drug Discov. Today 11, 355–362.
Schekman, R., 2013. How journals like Nature, Cell and Science are damaging science. Guardian, Available from: http://www.theguardian.com/commentisfree/2013/dec/09/how-journals-nature-science-cell-damage-science.
Schmidt, C.W., 2014. Research wranglers: initiative to improve reproducibility of study findings. Environ. Health Perspec. 122, A188–A191.
Schmidt, S., 2009. Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Rev. Gen. Psychol. 13, 90–100.
Scott, M., 2016. Everything you need to know about the Theranos saga so far. Wired. Available from: http://www.wired.com/2016/05/everything-need-know-theranos-saga-far/.
Scott, S., Kranz, J.E., Cole, J., Lincecum, J.M., Thompson, K., et al., 2008. Design, power, and interpretation of studies in the standard murine model of ALS. Amyotroph. Lateral Scler. 9, 4–15.
Scotter, E.L., Chen, H.J., Shaw, C.E., 2015. TDP-43 proteinopathy and ALS: insights into disease mechanisms and therapeutic targets. Neurotherapeutics 12, 352.
Sen, S., 2011. Francis Galton and regression to the mean. Significance (The Royal Statistical Society), pp. 124–126.
Sena, E.S., van der Worp, H.B., Bath, P.M.W., Howells, D.W., Macleod, M.R., 2010. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol. 8, e1000344.
Serikawa, K., 2015. Baseball, regression to the mean, and avoiding potential clinical trial biases. Available from: https://kyleserikawa.com/2015/05/18/baseball-regression-to-the-mean-and-avoiding-potential-clinical-trial-biases/.
Senapathy, K., 2016. No Andrew Wakefield, You're Not Being Censored And You Don't Deserve Due Process. Forbes.com Opinion. Available from: http://www.forbes.com/sites/kavinsenapathy/2016/03/28/no-andrew-wakefield-youre-not-being-censored-and-you-dont-deserve-due-process/#6715cbc0225d.
Shafer, S., 2011. Research Fraud in Anesthesia. American Society of Anesthesiologists Newsletter. Available from: http://www.asahq.org/resources/publications/newsletter-articles/2011/may2011/research-fraud-in-anesthesia.
Shen, C., Björk, B.-C., 2015. Predatory open access: a longitudinal study of article volumes and market characteristics. BMC Med. 13, 230.
Shimasaki, C., 2014. Biotechnology Entrepreneurship. Starting, Managing, and Leading Biotech Companies. Elsevier Academic, Waltham, MA.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22, 1359–1366.
Singhal, A., Leaman, R., Catlett, N., Lemberger, T., McEntyre, J., Polson, et al., 2016. Pressing needs of biomedical text mining in biocuration and beyond: opportunities and challenges. Database 2016.
Siva, N., 2010. Wakefield's first try. Slate. Available from: http://www.slate.com/articles/health_and_science/medical_examiner/2010/06/wakefields_first_try.html.
Smaldino, P.E., McElreath, R., 2016. The natural selection of bad science. R. Sci. Open Sci. 3, 160384.
Smith, R., 2006. Peer review: a flawed process at the heart of science and journals. J. R. Soc. Med. 99, 178–182.
Snyder, P.J., Mayes, L.C., Smith, W.E., 2015. The Management of Scientific Integrity within Academic Medical Centers. Academic Press, San Diego.
Song, F., Hooper, L., Loke, Y.K., 2013. Publication bias: what is it? How do we measure it? How do we avoid it? Open Access J. Clin. Trials 2013, 71–81.
Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, R., et al., 2014. Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods 11, 629–632.
Souder, L., 2011. The ethics of scholarly peer review: a review of the literature. Learned Pub. 24, 55–74.

Stapel, D.A., Lindenberg, S., 2011. Retracted: coping with chaos: how disordered contexts promote stereotyping and discrimination. Science 332, 251–253.
Steen, R.G., 2011a. Retractions in the scientific literature: do authors deliberately commit research fraud? J. Med. Ethics 37, 113–117.
Steen, R.G., 2011b. Retractions in the scientific literature: is the incidence of research fraud increasing? J. Med. Ethics 37, 249–253.
Steen, R.G., 2011c. Retractions in the medical literature: how many patients are put at risk by flawed research? J. Med. Ethics 37, 688–692.
Steen, R.G., Casadevall, A., Fang, F.C., 2013. Why has the number of scientific retractions increased? PLoS One 8, e68397.
Stemwedel, J.D., 2015. Is aggressive science reporting a human rights violation? Forbes. Available from: http://www.forbes.com/sites/janetstemwedel/2015/08/29/is-aggressive-science-reporting-a-human-rights-violation/#4fa718257488.
Stewart, W.W., Feder, N., 1987. The integrity of the scientific literature. Nature 325, 207–214.
Stone, A., 2005. The cheating culture. BusinessWeek Archives. Available from: http://www.bloomberg.com/bw/stories/2005-06-20/the-cheating-culture.
Stoye, E., 2015. Post publication peer review comes of age. Chemistry World. Available from: http://www.rsc.org/chemistryworld/2015/01/post-publication-peer-review-stap-comes-age.
Stroebe, W., Postmes, T., Spears, R., 2012. Scientific misconduct and the myth of self-correction in science. Persp. Psychol Sci. 7, 670–688.
Sung, N.S., Crowley, Jr., W.F., Genel, M., Salber, P., Sandy, L., et al., 2003. Central challenges facing the national clinical research enterprise. JAMA 289, 1278–1287.
Suresh, K.P., 2011. An overview of randomization techniques: an unbiased assessment of outcome in clinical research. J. Hum. Reprod. Sci. 4, 8–11.
Tarkan, L., 2016. Why Robert De Niro Promoted—then Pulled—Anti-Vaccine Documentary. Fortune. Available from: http://fortune.com/2016/03/29/robert-de-niro-anti-vaccine-documentary/.
ter Riet, G., Korevaar, D.A., Leenaars, M., Sterk, P.J., Van Noorden, C.J.F., et al., 2012. Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. PLoS One 7, e43404.
Tharyan, P., 2012. Criminals in the citadel and deceit all along the watchtower: irresponsibility, fraud, and complicity in the search for scientific truth. Mens Sana Monogr. 10, 158–180.
Thomas, J., 2010. Paranoia Strikes Deep: MMR Vaccine and Autism. Psychiatric Times. Available from: http://www.psychiatrictimes.com/autism/%E2%80%9Cparanoia-strikes-deep%E2%80%9D-mmr-vaccine-and-autism#sthash.PDAqrm2v.dpuf.
Torrente, A., Lukk, M., Xue, V., Parkinson, H., Rung, J., Brazma, A., 2016. Identification of cancer related genes using a comprehensive map of human gene expression. PLoS One 11 (6), e0157484.
Triggle, D.J., Miller, K.W., 2002. Doctoral education: another tragedy of the commons? Am. J. Pharm. Edu. 66, 287–294.
Triggle, C.R., Williams, M., 2015. Challenges in the biomedical research enterprise in the 21st century: antecedents in the writings of David Triggle. Biochem. Pharmacol. 98, 342–359.
Tsilidis, K.K., Panagiotou, O.A., Sena, E.S., Aretouli, E., Evangelou, E., Howells, D.W., et al., 2013. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biol. 11, e1001609.
van Dalen, H.P., Henkens, K., 2012. Intended and unintended consequences of a publish-or-perish culture: a worldwide survey. J. Am. Soc. Inform. Sci. Technol. 63, 1282–1293.
Van der Staay, F.J., Steckler, T., 2002. The fallacy of behavioral phenotyping without standardization. Genes Brain Behav. 1, 9–13.
van der Vet, P.E., Nijveen, H., 2016. Propagation of errors in citation networks: a study involving the entire citation network of a widely cited paper published in, and later retracted from, the journal Nature. Res. Integr. Peer Rev. 1, 3.
van der Worp, H.B., van Gijn, J., 2007. Clinical practice. Acute ischemic stroke. N. Engl. J. Med. 357, 572–579.
van der Worp, H.B., Howells, D.W., Sena, E.S., Porritt, M.J., Rewell, S., O'Collins, V., et al., 2010. Can animal models of disease reliably inform human studies? PLoS Med. 7, e1000245.
Vanden Berghe, T., Hulpiau, P., Martens, L., Vandenbroucke, R.E., Van Wonterghem, E., Perry, S.W., et al., 2015. Passenger mutations confound interpretation of all genetically modified congenic mice. Immunity 42, 200–209.
Van Noorden, R., 2011. Science publishing: the trouble with retractions. Nature 478, 26–28.
Van Noorden, R., 2014. Parasite test shows where validation studies can go wrong. Nature. Available from: http://www.nature.com/news/parasite-test-shows-where-validation-studies-can-go-wrong-1.16527.

Vandamme, T.F., 2015. Rodent models for human diseases. Eur. J. Pharmacol. 759, 84–89.
Vasagar, J., 2001. Rise of the wealthy Oxford scientists. Guardian. Available from: https://www.theguardian.com/uk/2001/apr/21/highereducation.education.
Vasilevsky, N.A., Brush, M.H., Paddock, H., Ponting, L., Tripathy, S.J., LaRocca, G.M., Haendel, M., 2013. On the reproducibility of science: unique identification of research resources in the biomedical literature. Peer J. 1, e148.
Vatner, S.F., 2016. Why so few new cardiovascular drugs translate to the clinics. Circ. Res. 119, 714–717.
Verfaellie, M., McGwin, J., 2011. The case of Diederik Stapel. Psychological Science Agenda, American Psychological Association. Available from: http://www.apa.org/science/about/psa/2011/12/diederik-stapel.aspx.
Wade, N., 2010. Harvard Finds Scientist Guilty of Misconduct. New York Times. Available from: http://www.nytimes.com/2010/08/21/education/21harvard.html?_r=0.
Wadman, M., 2013. NIH mulls rules for validating key results. Nature 500, 14–16.
Wakefield, A.J., Pittilo, R.M., Sim, R., Cosby, S.L., Stephenson, J.R., Dhillon, A.P., Pounder, R.E., 1993. Evidence of persistent measles virus infection in Crohn's disease. J. Med. Virol. 39, 345–353.
Wakefield, A.J., Murch, S.H., Anthony, A., Linnell, J., Casson, D.M., Malik, M., et al., 1998. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 351, 637–641.
Ware, M., Mabe, M., 2015. STM Report, fourth ed. Available from: http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf.
Weaver, C., 2017. Theranos secretly bought outside lab gear and ran fake tests, court filings allege. Wall Street J. Available from: https://www.wsj.com/articles/theranos-secretly-bought-outside-lab-gear-ran-fake-tests-court-filings-1492794470.
Wehling, M., 2009. Assessing the translatability of drug projects: what needs to be scored to predict success? Nat. Rev. Drug Discov. 8, 541–546.
Weichenberger, C.X., Pozharski, E., Rupp, B., 2017. Twilight reloaded: the peptide experience. Acta Cryst. D73, 211–222.
Wells, J.A., 2008. Final Report: Observing and Reporting Suspected Misconduct in Biomedical Research Gallup/Office of Research Integrity. Available from: http://ori.hhs.gov/sites/default/files/gallup_finalreport.pdf.
Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., et al., 2014. The NHGRI GWAS catalog, a curated resource for SNP-trait associations. Nucl. Acids Res. 42, D1001–D1006.
Wendler, A., Wehling, M., 2012. Translatability scoring in drug development: eight case studies. J. Transl. Med. 10, 39.
Wilbanks, J., Friend, S.H., 2016. First, design for data sharing. Nat. Biotech. 34, 377–379.
Wilson, W.A., 2016. Scientific regress. First Things. Available from: http://www.firstthings.com/article/2016/05/scientific-regress.
Wise, J., 2013. Boldt: the great pretender. BMJ 346, f1738.
Witkoski, T., 2014. From the archives of scientific fraud—Diederik Stapel. Psychology Gone Wrong. The Dark Sides of Science and Therapy. Available from: https://forbiddenpsychology.wordpress.com/2014/06/26/from-the-archives-of-scientific-fraud-diederik-stapel/.
Woodward, J., Goodstein, D., 1996. Conduct, misconduct, and the structure of science. Amer Sci 84, 468–478.
Young, N.S., Ioannidis, J.P.A., Al-Ubaydi, O., 2008. Why current publication practices may distort science. PLoS Med. 5, 1418–1422.
Yong, E., 2012. Uncertainty shrouds psychologist's resignation. Nature. Available from: http://www.nature.com/news/uncertainty-shrouds-psychologist-s-resignation-1.10968.
Yong, E., Ledford, H., Van Noorden, R., 2014. Research ethics: 3 ways to blow the whistle. Nature 503, 454–457.
Zoghabi, H.Y., 2013. The basics of translation. Science 339, 250.