J. Dairy Sci. 101:1–23
https://doi.org/10.3168/jds.2017-13978

Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences

Nora M. Bello*†‡[1] and David G. Renter‡§
*Department of Animal Science, University of Wisconsin, Madison, WI 53706
†Department of Statistics,
‡Center for Outcomes Research and Epidemiology, and
§Department of Diagnostic Medicine and Pathobiology, Kansas State University, Manhattan 66506

ABSTRACT

Reproducible results define the very core of scientific integrity in modern research. Yet, legitimate concerns have been raised about the reproducibility of research findings, with important implications for the advancement of science and for public support. With statistical practice increasingly becoming an essential component of research efforts across the sciences, this review article highlights the compelling role of statistics in ensuring that research findings in the animal sciences are reproducible—in other words, able to withstand close interrogation and independent validation. Statistics set a formal framework and a practical toolbox that, when properly implemented, can recover signal from noisy data. Yet, misconceptions and misuse of statistics are recognized as top contributing factors to the reproducibility crisis. In this article, we revisit foundational statistical concepts relevant to reproducible research in the context of the animal sciences, raise awareness on common statistical misuse undermining it, and outline recommendations for statistical practice. Specifically, we emphasize a keen understanding of the data generation process throughout the research endeavor, from thoughtful experimental design and randomization, through rigorous data analysis and inference, to careful wording in communicating research results to peer scientists and society in general. We provide a detailed discussion of core concepts in experimental design, including data architecture, experimental replication, and subsampling, and elaborate on practical implications for proper elicitation of the scope of reach of research findings. For data analysis, we emphasize proper implementation of mixed models, in terms of both distributional assumptions and specification of fixed and random effects to explicitly recognize multilevel data architecture. This is critical to ensure that experimental error for treatments of interest is properly recognized and inference is correctly calibrated. Inferential misinterpretations associated with use of P-values, both significant and not, are clarified, and problems associated with error inflation due to multiple comparisons and selective reporting are illustrated. Overall, we advocate for a responsible practice of statistics in the animal sciences, with an emphasis on continuing quantitative education and interdisciplinary collaboration between animal scientists and statisticians to maximize reproducibility of research findings.
Key words: experimental replication, model specification, statistical errors, multiple testing

INTRODUCTION

Reproducible results define the very core of scientific integrity in modern research. Yet, legitimate questions have been recently raised about the robustness and reliability of new scientific knowledge (Ioannidis, 2005) based on a reported inability to reproduce research results (Ioannidis et al., 2009; Begley and Ellis, 2012; Begley and Ioannidis, 2015). Collectively, these concerns have come to be known as research reproducibility (**RR**) issues and seem to be quite pervasive across scientific disciplines (Begley and Ioannidis, 2015; Nuzzo, 2015; Baker, 2016).

To illustrate the magnitude of the RR crisis, consider the following evidence. In 2009, Ioannidis et al. (2009) reevaluated 18 microarray-based gene expression studies, of which only 2 were fully reproduced. Soon afterward, in 2012, scientists reported replicating only 6 out of 53 landmark studies on preclinical cancer efforts for drug development (Begley and Ellis, 2012). Meanwhile, a large collaborative project attempted replication of 100 experiments published in high-ranking psychology journals and succeeded only one-third to one-half of the time, depending on the criteria (Open Science Collaboration, 2015). A recent compilation of biomedical studies showed that approximately 75 to 90% of results in preclinical research, and as much as 85% of research

Received October 11, 2017.
Accepted March 8, 2018.
[1] Corresponding author: nbello@ksu.edu

results across the biomedical spectrum, were deemed irreproducible (Begley and Ioannidis, 2015). A survey of *Nature* readers revealed that as many as 70% of researchers have tried and failed to reproduce another scientist's research results and over 50% have failed to reproduce their own results (Baker, 2016). Interest in RR is further evidenced in special dedicated issues in high-profile journals, including *Science* (http://www.sciencemag.org/site/special/data-rep/) and *Nature* (http://www.nature.com/news/reproducibility-1.17552). Even lay publications (Young and Karr, 2011) and mainstream newspapers (Shaywitz, 2009; Naik, 2011) have reported on the subject of RR. Taken together, the extent and scope of the reproducibility crisis is broad, encompassing many basic scientific fields as well as more applied disciplines (reviewed by Begley and Ioannidis, 2015; Baker, 2016). In the animal sciences, the full extent of the RR crisis has not yet been characterized in detail, though the issue has been discussed previously (Bello et al., 2016; Tempelman, 2016).

Admittedly, the multifaceted and increasingly complex nature of many of the research problems currently at the forefront of science adds to the complications of the RR issue. Specific to agriculture, one may consider the daunting challenge of meeting demands for a safe and secure food supply for an exponentially growing world population under competing demands for environmental sustainability and limited natural resources in a changing climate. Meanwhile, the research process is becoming more and more quantitative across scientific domains. This seems largely attributed to increasingly large data sets and new data types, which in turn call for more and more sophisticated quantitative methods that often require specialized expertise. Overall, this state of affairs argues for dynamic multidisciplinary integration of scientific disciplines, for which a common, presumably quantitative, language is imperative. In this context, statistical practice is increasingly becoming a critical component of research efforts across scientific disciplines, including the animal sciences.

By its very interdisciplinary nature, statistics is uniquely poised to help address the RR crisis. Indeed, statistics, as a discipline, provides both a common quantitative language to help bridge scientific domains and a sophisticated formal framework that, when properly implemented, allows for the recovery of signal from the inherently noisy nature of data. In this review, we focus on the role of statistical practice in ensuring RR in animal health and production. Often defined as the science of learning from data, the statistical sciences offer both conceptual infrastructure and practical tools to deal with the many sources of variability naturally embedded in complex systems. However, for all its sophistication, statistics is not a silver bullet for dealing with noisy data; the most one can expect is proof beyond a reasonable doubt, though hardly 100% certainty. As famously discussed by statistician George Box (1919–2013), "all models are wrong...," as, even in best-case scenarios, statistical models are often simplistically naive and only provide rough approximations of a much more complex reality. Nevertheless, so continues Box's quote, when properly implemented, "... some [models] can be useful," mainly by introducing a level of robustness and rigor within which the process of learning from data can be tackled more objectively. Particularly with large data sets, foundational statistical concepts take renewed relevance, as flaws in experimental design and bias of inference and prediction cannot be overcome by large sample sizes, as conspicuously illustrated by the so-called parable of Google Flu (Lazer et al., 2014). While there undoubtedly is valuable signal to be extracted from large data sets, "big" is not always better. That is, more data does not necessarily produce more information; rather, it may muddy the waters with irrelevant information. We argue that it is precisely in the context of big data that the ideas of sound experimental design and well-implemented data analysis may turn out to be as, or more, important than ever in the research process. To this end, the 2016 joint meeting of the American Dairy Science Association and the American Society of Animal Sciences had a special session on "Big Data in Animal Science: Uses for Models, Statistics and Meta-Approaches" (http://www.jtmtg.org/JAM/2016/abstracts/612.pdf).

Misunderstanding and misuse of statistical concepts are contributing factors to the RR crisis, both due to their nature (Ioannidis, 2005; Nuzzo, 2015) and to their widespread ubiquity across the sciences (Reinhart, 2015). Though with fewer documented examples, production agriculture (Sargeant et al., 2009, 2011; Kramer et al., 2016), and more specifically the animal sciences (Tempelman, 2009; Bello et al., 2016), are no exception. Therefore, our specific objectives in this review were (1) to reexamine foundational statistical concepts relevant to RR in the context of the animal sciences, (2) to educate the animal science community about common statistical misuses that undermine RR, and (3) to outline guidelines and introduce tools for statistical practice that maximize legitimacy and credibility of new scientific knowledge.

Before proceeding, a couple of disclaimers are in order. First, we emphasize foundational statistical principles, both for experimental design and for data analysis and interpretation. Admittedly, most of these principles are hardly novel in and of themselves; yet, they can be highly nuanced and have implications not immediately obvious to the animal scientist in the evolving land-