# Never Waste a Good Crisis: Confronting Reproducibility in Translational Research

Daniel J. Drucker[1],*
[1]Department of Medicine, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, ON M5G 1X5, Canada
*Correspondence: drucker@lunenfeld.ca

http://dx.doi.org/10.1016/j.cmet.2016.08.006

The lack of reproducibility of preclinical experimentation has implications for sustaining trust in and ensuring the viability and funding of the academic research enterprise. Here I identify problematic behaviors and practices and suggest solutions to enhance reproducibility in translational research.

As I contemplated the content of my lecture at a recent Keystone symposium, the potential topics to be addressed were tantalizing. The theme of this Keystone meeting, ''New Therapeutics for Diabetes and Obesity,'' was nicely aligned with the interests of my lab and our studies examining the therapeutic potential of gut hormones. On the program was a litany of leading scientists discussing the most exciting advances in metabolic disease research, encompassing brown fat, central nervous system control of glucose and body weight, advances in islet biology, stem cells, peptide therapeutics, mouse and human genetics, immunotherapy, and neuromodulation. There were even lectures devoted to discussing existing and novel preclinical animal models widely used to test promising pathways and compounds for efficacy in treating experimental diabetes and obesity. There has rarely been a more exciting time to unravel the molecular mechanisms and pathways controlling energy intake and assimilation, and the abnormalities in these pathways that predispose us to the development of diabetes and obesity.

In the back of my mind however, was the ever-present nagging voice of sobering reality. The voice reminds me, a clinician scientist, that the vast majority of genes and proteins and pathways and targets that provide impressive and exciting results in preclinical studies on a daily basis (and a justifiable number of exciting high-profile widely publicized publications) usually do not survive the difficult path toward rigorous target validation and ultimate clinical development. Although the joy of an exciting result in preclinical studies need not be extinguished by the stark reality that many findings will simply not be reproducible in animals, let alone translatable in human studies, we have developed a culture of hype and exaggerated expectations that often fall far short of the promises made. These reproducibility challenges are not confined to the study of metabolism and are independent of the much less common, but equally worrisome, issue of scientific misconduct, which continues, like the Lernean Hydra, to raise its ugly head on a daily basis no matter how hard the community tries to extinguish egregious scientific behavior.

Reproducibility issues in basic science are now being debated regularly. Published surveys from industry colleagues routinely highlight difficulties in validation or reproduction of major research findings from academic laboratories (Prinz et al., 2011). The National Institutes of Health leadership has identified a number of remediable problems worthy of correction that will address experimental design, pitfalls in animal experimentation, use of appropriate statistics, transparency of methodology, and over-interpretation of data (Collins and Tabak, 2014). Indeed, agencies within the NIH have convened conferences to discuss reproducibility challenges plaguing preclinical research providing constructive guidance and recommended reporting standards designed to enhance transparency and reproducibility (Landis et al., 2012). These recommendations include calls for random assignment of animals, blinding of preclinical treatment groups, more rigorous sample and effect size calculations, and formal rules for handling of data involving outliers, pre-specified primary and secondary endpoints, and replication of key experimental findings (Landis et al., 2012). While laudable, these recommendations have not been formally adopted by journals, and reading the metabolism literature suggests that the majority of laboratories do not strictly adhere to these ''best practices.''

Herein, I discuss challenges and issues that might account for some of the chasm between our astounding collective success in explaining, treating, and often vanquishing metabolic disease in preclinical studies, and our meager success rate in moving most of these discoveries from bench to bedside. The recurring spectacular preclinical discovery paradigm is often highlighted at Keystone or related meetings, prompting me to consider our collective expensive, often disappointing, inability to reproduce and translate the majority of early-stage exciting research into therapeutic interventions with clinical utility. Even attempting to precisely define reproducibility can be contentious, as the term embodies different concepts, accepted ranges of variability, and varying criteria that differ across fields and scientific communities (Goodman et al., 2016). Using examples from our own area of gut hormone research, I illustrate anecdotally some commonly encountered challenges and pitfalls in the design and interpretation of experimental data. Finally, in an attempt to stimulate discussion, I provide examples of and suggestions for fueling ongoing efforts to improve and standardize preclinical research, so our own community may find greater success in the pursuit of preclinical reproducibility and, ultimately, enhanced bench-to-bedside translatability.

## Execution and Reporting of Clinical versus Preclinical Studies

Human clinical trials are often carried out using a randomized double-blinded design, and outliers, or suboptimal responders, are not discarded from the analysis. It is expected that clinical researchers will ideally account for and report on every single study subject screened, and ultimately enrolled in a clinical trial, even if subjects drop out or move away. Non-responders are not simply discarded from the trial results, and there are statistical methods employed to account for study subjects who may not complete the entire study. Both efficacy and safety are critically scrutinized, and the primary and secondary outcomes must be carefully stated in advance, thus minimizing extensive post hoc number crunching to find statistically significant unexpected outcomes. Moreover, many large clinical trials study large numbers of genetically diverse subjects, from different regions of the world, both male and female, often including a wide range of ages. While it is certainly true that clinical trials reporting negative results are more slowly reported, and sometimes not at all, efforts to improve reporting, such as the initiative embodied within alltrials.net, are likely to produce further improvements in comprehensive reporting. It would be viewed as completely unacceptable for an academic clinical trials unit to carry out dozens of clinical studies in human subjects and only report the most promising results from studies that produced a favorable outcome.

Contrast this situation with current norms and expectations for preclinical studies and research in animals. In many discovery-focused basic science laboratories, we seek to understand disease pathophysiology, often honing in on a small number of genes, proteins, and pathways that receive intensive scrutiny. The research is frequently exploratory and open-ended, with large numbers of variables and endpoints simultaneously quantified. It is common practice to perform dozens, if not hundreds, of experiments in cells, mice, and rats, yet often only a small subset of the data is reported and made available for scrutiny. Not surprisingly, the majority of published manuscripts contain the most promising and exciting results that ‘‘worked the best’’ or generated data and yielded mechanisms consistent with the prevalent hypothesis. What happens to the dozens of experiments that ‘‘did not work,’’ a euphemism meaning that the results of the studies did not turn out the way the scientists wanted and/or did not support the story being assembled in the laboratory? Rather than accepting that our theories may be incorrect or not important, or that our favorite molecule may not produce robust and reproducible actions in multiple models, we frequently soldier on, trying different time points, concentrations, conditions, and animal models, until we get just the ‘‘right result,’’ which ends up in a figure in our paper. A majority of negative, contradictory, or divergent results may never see the light of the day and are not reported. How would scientists (and journal reviewers and editors) react to a proposal mandating accounting for and reporting of all results in all animal studies, such as a formal standardized structured report (Figure 1) perhaps included as a mandatory online appendix, accessible to the reviewer and, ultimately, the reader? This transparency would likely prove sobering and for some would be an unwelcome obligation, yet it might frame the more promising results presented in a more realistic light and broader context.

There is a great deal to be said for exploratory discovery research, uncovering findings and mechanisms that account for some, but not all, of the pathophysiology in a particular disease model. However, our temptation to generalize and elevate the significance of our positive results likely contributes to future challenges that arise when other scientists, using slightly different methods, doses, conditions, and animals, cannot reproduce the results we publish.

## How Reproducible Is Basic Science Research?

Many scientists might scoff with indignation if someone questions reproducibility of their own research; after all, this ‘‘reproducibility issue’’ is usually someone else’s problem. Nevertheless, some honest individuals have thoughtfully discussed challenges inherent in reproducing observations within their own lab, let alone across laboratories (Woods and Begg, 2015). The available evidence from multiple fields clearly indicates that we have a systemic reproducibility problem in basic science research. Below, I review some of the contemporary issues that contribute to reproducibility challenges within preclinical metabolic research. It is likely that many of the problems to be confronted are relevant to many others outside the field of metabolism, and the issues raised may resonate with a broader community of basic scientists.

### *Cell Lines*

Considerable debate has focused on the identification and reliability of cell lines and, while progress has been made in this area, the problem continues to fester. As a postdoctoral fellow in the mid-1980s, I was excited to have isolated, with colleagues, a new human glucagon-producing cell line. We were convinced this would be an invaluable reagent for study of human glucagon biosynthesis and secretion, and we had assembled a good many figures for our envisioned paper. Like many things in life, what seemed too good to be true actually was; analysis of genomic DNA from my ‘‘human cells’’ revealed the presence of repeated DNA sequences from both human and rat DNA. It turned out that our ‘‘human glucagonoma’’ cell line was likely a mixture of HeLa cells and our new RIN1056A glucagon-producing cell line, and the party was over. Several years later, we also discovered mycoplasma contamination of our hamster glucagon-producing cell line and wasted several valuable months redoing key experiments after re-deriving ‘‘mycoplasma-free’’ InR1G9 hamster glucagonoma cells. In hindsight, it was a valuable learning experience to identify, early on, the pitfalls of using incompletely characterized or infected cell lines for basic science studies.

Surprisingly, however, although considerable effort has been devoted to recognition of the problems associated with cell line identification and verification (Freedman et al., 2015), there is scant evidence that scientists have routinely adopted these guidelines to ensure the fidelity and rigor of their own cell line research. The origin and sourcing of cell lines is often inadequately described in publications, further challenging efforts to reproduce published data. While differences in source of cell line, passage number, cell density, cell culture conditions, and experimental technique may logically account for some degree of variability, many publications provide inadequate details that

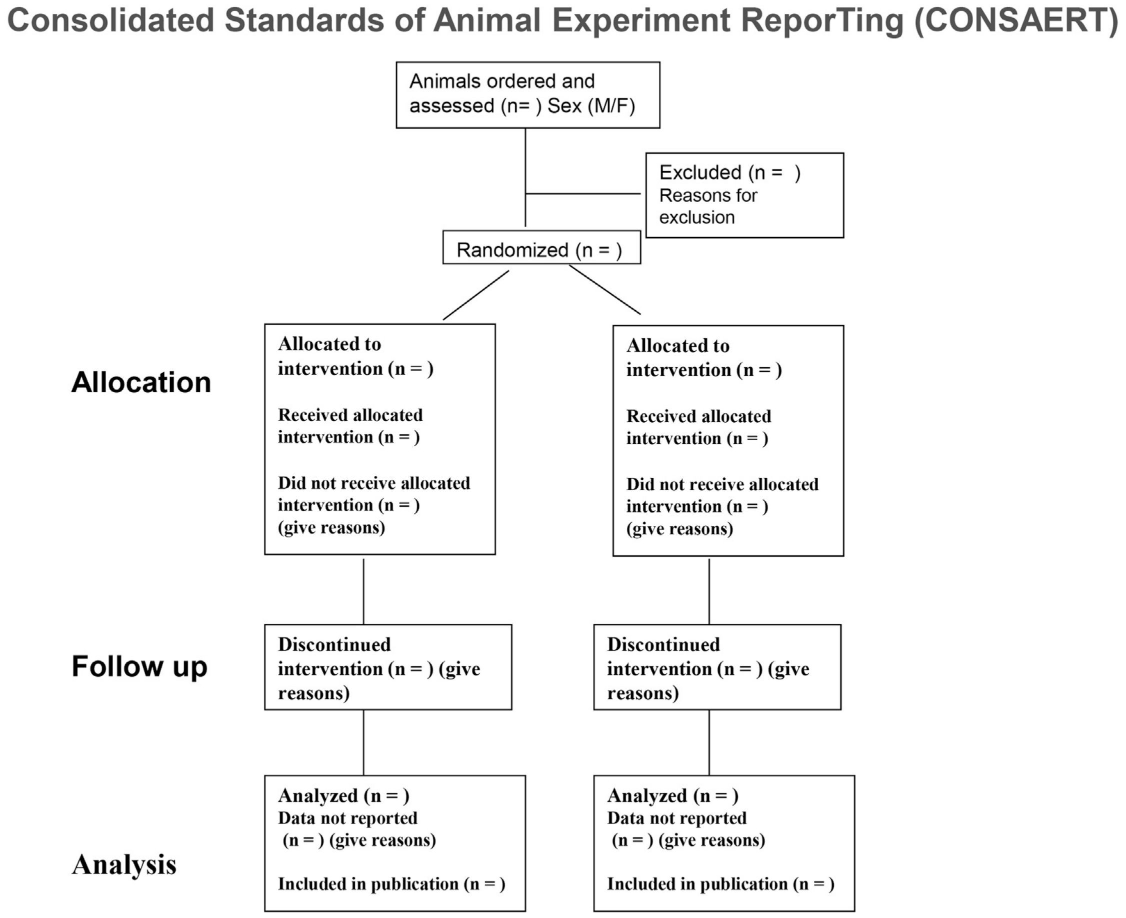
Consolidated Standards of Animal Experiment ReporTing (CONSAERT)
Animals ordered and assessed (n= ) Sex (M/F)
Excluded (n = ) Reasons for exclusion
Randomized (n = )
Allocation
Allocated to intervention (n = )
Received allocated intervention (n = )
Did not receive allocated intervention (n = ) (give reasons)
Allocated to intervention (n = )
Received allocated intervention (n = )
Did not receive allocated intervention (n = ) (give reasons)
Follow up
Discontinued intervention (n = ) (give reasons)
Discontinued intervention (n = ) (give reasons)
Analysis
Analyzed (n = ) Data not reported (n = ) (give reasons)
Included in publication (n = )
Analyzed (n = ) Data not reported (n = ) (give reasons)
Included in publication (n = )

**Figure 1. Proposed Template for Reporting Animal Use and Analysis in Preclinical Studies**
Designated the Consolidated Standards of Animal Experiment ReporTing (CONSAERT) flow diagram.

preclude careful reproduction of cell line experiments. Indeed, the trend in many journals is toward minimizing the length of methods sections. The regular use of simple quality control techniques to verify cell line identity and potential contamination would greatly enhance the validity of cell line data, yet many institutions, granting agencies, and journals do not regularly insist on obligatory detailed cell line reporting (Freedman et al., 2015). With the emerging use of stem cell-derived cells for the study of islet biology, it seems likely that very precise detailed disclosure of the exact composition of cell culture media and all essential exogenous additives and growth factors, coupled with specification of the gender and age of origin as well as extensive molecular characterization and footprinting criteria, will be needed to ensure replication of stem cell-derived cells within and across laboratories.

### *Antibodies-Quicksand for the Non-curious*

Equally vexatious is the ongoing crisis promulgated by use of antibodies that have not been properly validated and, as a result, generate irreproducible or incorrect data due to lack of sensitivity and/or problems with specificity. This challenge extends to all fields of research that use antibodies, and every researcher has their own story with ''problematic antibodies.'' In the incretin field, there are dozens of published papers using commercial antibodies employed to detect the GLP-1 receptor; our own laboratory experience, regrettably, is that most of these antibodies do not detect the GLP-1 receptor. The use of non-specific antibodies, together with studies employing small numbers of animals and inadequate controls for animal and histology experiments, has led to significant misconceptions in the incretin field and supported, in part, the imbroglio surrounding GLP-1R expression in normal and neoplastic pancreatic cells (Drucker, 2013).

Following several publications and editorials (Gore, 2013; Panjwani et al., 2013; Pyke et al., 2014; Pyke and Knudsen, 2013) highlighting the major problems and inadequacies associated with the use of flawed GLP-1R antisera, one might have hoped that dissemination of this information through publications and discussions at meetings might lead to elimination of the problem. Indeed, many editorials continue to highlight the importance of extensive antibody validation. Sadly, although our paper describing problems with the sensitivity and specificity of GLP-1R antisera appeared online in November 2012, I estimate that about every other week I still read another new publication reporting data using suspect or incompletely characterized GLP-1R antisera (Panjwani et al., 2013). What does this say about the thoroughness and credibility of our community of reviewers, editors, and scientific colleagues? Not surprisingly, we also encounter substantial problems with sensitivity and specificity of antisera widely used to detect the GLP-2 receptor (Drucker and Yusta, 2014), and our unpublished studies raise similar issues in regard to the specificity of antibodies for the glucagon and GIP receptors.

The pervasive problem associated with inadequate antibodies has received widespread attention; however, both experienced and junior investigators still fail to routinely characterize antibodies used in their laboratories. A survey of over 500 scientists revealed that less than half of junior investigators (<5 years out) reported validating antibodies. This data speaks to our ongoing need as a research community to properly educate our trainees in the appropriate characterization, validation, and use of antibodies for research purposes (Freedman et al., 2016). How can the antibody problem be solved? The increasingly frank public discussion of problematic antibodies, coupled with publicly available databases documenting inadequate or appropriate antibody validation (Baker, 2015), represents important steps in energizing the community to pay more attention to the quality of antibodies. The use of CRISPR technology enables rapid generation of ‘‘knockout’’ cells, facilitating assessment of antibody specificity. While some scientists and a few antibody companies have enjoined discussions to more rigorously characterize antibody sensitivity and specificity (Bradbury and Plückthun, 2015), including calls for use of DNA sequence verification and use of strictly recombinant antibodies, progress has been slow. Nevertheless, there seems to be a clear commercial opportunity for the next-generation antibody company that works to develop a reputation for excellence in antibody characterization, a reputation not currently enjoyed by any of the companies presently operating in the space.

## Animal Models for Metabolism Research

Many scientists study the pathophysiology of diabetes using animal models, with a view to development of novel therapeutic agents. In some instances, we use these animal models to interrogate mechanisms and identify key genes and proteins underlying islet dysfunction, disordered hepatic glucose production, or insulin resistance arising through disturbances in signaling within muscle, adipose tissue, brain, and other organs. Alternatively, we may already have developed promising new therapeutic agents for the treatment of diabetes or obesity and seek to test their efficacy in representative animal models. Surprisingly, despite broad awareness that the majority of risk for development of human diabetes arises through small contributions from dozens of genes with modest effect sizes (Fuchsberger et al., 2016), the diabetes research community continues to heavily utilize predominantly monogenic models of disease, including the Akita, ob/ob, and db/db mouse, and corresponding rat models such as the Zucker fatty and Zucker diabetic fatty rat. While the seminal metabolic importance of the leptin signaling pathway is beyond dispute, the genes encoding leptin or the leptin receptor are not associated with the risk of developing type 2 diabetes (T2D) in human population studies. While these rodent models recapitulate some, if not many, of the features encountered in human subjects with diabetes, obesity, and insulin resistance, they are likely suboptimal models with which to study potential therapies for T2D. Most of these animal models develop diabetes and obesity at accelerated rates, some associated with rapid development of β cell failure, quite unlike the indolent slowly progressive natural history of human T2D (Wang et al., 2014). Although administration of leptin is strikingly effective in ameliorating diabetes in leptin-sensitive mice and rats, the same cannot be said for the efficacy of leptin in human subjects with T2D or in insulin-treated subjects with type 1 diabetes (T1D) (Vasandani et al., 2016). Similar challenges surround the extensive use of the *Ins2*$^{Akita}$ mouse for studies of β cell failure and diabetic nephropathy independent of obesity and insulin resistance. While the *Ins2*$^{Akita}$ mouse is an excellent animal model for analysis of endoplasmic reticulum stress and β cell failure, mutations within the human insulin gene do not make significant contributions to the genetic risk for development of T2D. Although much more expensive and time consuming, several months of high-fat feeding in diabetes-prone mice or rats is more likely to recapitulate many of the features and natural history evident in human subjects with slowly progressive weight gain who ultimately develop T2D.

Equally problematic may be the extensive reliance on male mice and rats, as development of hyperglycemia and diabetes emerges less frequently in female mice in widely used strains. Notwithstanding directives from NIH and other granting agencies to ensure balanced use of both male and female animals and cell lines in preclinical studies (Clayton and Collins, 2014), the impact and ultimate success of these initiatives remains unclear. Given the widespread epidemic of diabetes and obesity in women, how well do findings made predominantly in studies of male mice and rats with diabetes predict efficacy in future translational studies of women? Although the use of non-human primates for diabetes research might yield data more easily reproduced in human clinical studies (Harwood et al., 2012), the limitations and expense inherent in these studies suggest that they are likely to be valuable for confirmatory, rather than exploratory or discovery, experiments.

## How Reproducible Are Mouse Experiments?

The reproducibility of animal data may reflect challenges with experimental design, small numbers, inadequate statistical analyses, or failure to communicate sufficient information to allow careful repetition of the experiments (Kilkenny et al., 2009). Differences among animal facilities including noise, bedding, diet, water supply, circadian rhythms, light/dark cycles, and gut microbial populations can greatly influence murine metabolic phenotypes. Indeed, even within experienced consortia dedicated to the detailed and careful standardization of mouse phenotyping across sites, considerable variation in phenotyping may still exist within and between centers (Hrabě de Angelis et al., 2015).

As someone who spends a great deal of time studying the function of the gastrointestinal tract, variability in phenotypic responses in different mouse experiments is not a trivial issue. Our laboratory has repeatedly encountered scenarios, often involving studies of gut inflammation, where we cannot precisely reproduce our own phenotypes or those reported by others often in papers published in very good journals. Whether it is in the extent of local and systemic inflammation induced by varying doses of lipopolysaccharide, the timing and magnitude of the development of inflammatory bowel disease in the *Il10*$^{-/-}$ mouse, the analysis of intestinal permeability, barrier function, and systemic inflammation following gut injury or high fat feeding, or the evanescent nature of gut inflammation that may be inconsistent from one study to another within our own laboratory, quantitative reproducibility is a recurring challenge for some of our experiments. We experienced these reproducibility challenges when we moved our laboratory across the street from the Toronto General Hospital to the Mount Sinai Hospital about 10 years ago. After re-deriving mouse lines and reanalyzing several of our most exciting gut phenotypes, we were stunned and disappointed to note that a few of our most exciting observations made in one mouse facility had simply failed to transfer and were no longer evident when we moved to a new animal facility across the street.

While it is widely recognized that germ-free mice exhibit profound metabolic differences compared to genetically identical, yet conventionally raised, animals (Claus et al., 2008), the impact of subtle changes in gut microbial populations may not always be accounted for when interpreting phenotypes. Not all laboratories have the intellectual curiosity and resources required to untangle the complex relationships between genetic strains, animal facility environments, varying microbial populations and specific diets, interactions that may produce important differences in key metabolic phenotypes within the same mouse line (Ussar et al., 2015). The importance of intestinal microbial dysbiosis is a research area potentially relevant to the pathophysiology and treatment of murine diabetes, obesity, and insulin resistance. Nevertheless, whether many exciting observations implying causality for dysbiosis can be reproducibly translated in humans with metabolic disorders remains uncertain.

## Phenotypes Arising in Mouse Knockouts: Caveat Emptor

The use of genetically modified mice, while tremendously important and often exciting, presents its own specific set of issues. Most scientists recognize that germline knockouts may present challenges in interpretation and translation due in part to the potential for developmental adaptation and physiological compensation secondary to loss of a key gene and protein from the earliest stages of development. Nevertheless, one reads every week of a potential new target for drug development in diabetes or obesity (and other disease areas), based in part on an exciting set of phenotypes arising in knockout mice. In our own field of gut hormone biology, the *Glp1r*$^{-/-}$ mouse with germline inactivation of the *Glp1r* exhibits resistance to diet-induced obesity, increased energy expenditure, enhanced insulin sensitivity, and preservation of glucose tolerance on different genetic backgrounds after high-fat feeding (Ayala et al., 2010; Hansotia et al., 2007; Scrocchi and Drucker, 1998). Had these phenotypes been published as part of the first early descriptions of GLP-1 action, no doubt we would all be found years later commiserating on the failure of GLP-1R antagonists to reduce body weight and lower glucose in human subjects with obesity and metabolic syndrome.

An important solution to the problem of germline gene inactivation was heralded by the introduction of tissue-specific knockout mice, using cell and tissue-specific expression of Cre recombinase to produce more selective inactivation of genes, often in a conditional manner. While this technology has diminished many of the problems with phenotypes arising secondarily to developmental compensation in multiple tissues, it comes with its own set of challenges. Cre was initially viewed as a highly active enzyme with minimal toxicity; however, it is difficult to find a cell type that is not impacted in some way by high-level Cre expression. Cardiovascular scientists are well aware of the cardiomyopathy associated with the targeting of Cre to cardiomyocytes, a problem that remains a major challenge in this field (Bersell et al., 2013; Koitabashi et al., 2009). The widespread use of RIP-Cre mice to inactivate genes in β cells has also been plagued by multiple problems, including the potential for Cre expression to impair insulin secretion (Lee et al., 2006; Magnuson and Osipovich, 2013). More recently, the unanticipated biological actions emanating from expression of a human growth hormone passenger gene in many of the Cre driver lines targeting β cells (Brouwers et al., 2014) further confounds interpretation of published data. Additionally, some insulin promoter-cre lines are also expressed in regions of the brain that control metabolism, further complicating interpretation of phenotypes (Magnuson and Osipovich, 2013; Wicksteed et al., 2010). Cre toxicity and ensuing DNA damage may also become more evident in proliferating or apoptotic cells, conditions common in studies of β cell biology (Schmidt-Supprian and Rajewsky, 2007). Hence, the β cell field is faced with the disquieting realization that some of the observations contained within dozens of papers published using elegant genetic technology to produce β cell knockouts may in fact contribute to results and interpretations that may be incorrect.

Although the toxicity emanating from Cre expression in numerous cell types is well known and easy to uncover in the published literature, many mouse studies do not routinely employ Cre-only controls, further complicating interpretation of some of the interesting phenotypes reported. Furthermore, careful systematic characterization of Cre expression from established Cre driver lines has revealed unexpected Cre activity (and corresponding genetic inactivation) in multiple cellular domains outside the predicted classical cell or tissue type targeted by the chosen promoter. Cre expression can also vary within littermates and be influenced by maternal or paternal origin effects (Heffner et al., 2012). Most scientists do not go hunting systematically for these potential problems, and it is likely that many of these issues go unrecognized, particularly in studies carried out without Cre-expressing littermate controls.

An alternative approach to restricting Cre toxicity and avoiding issues arising from inactivation of genes during development is the use of conditional knockout technology. While attractive with many advantages, this technology does not eliminate the potential for ‘‘off target’’ effects to complicate interpretation of phenotypes. It should not be surprising that tamoxifen, a potent steroid hormone, exerts biological activity itself in many cell types, including induction of apoptosis and metaplasia in the gastric epithelium (Huh et al., 2012), and lipoatrophy followed by de novo adipogenesis, manifest in an adipose tissue depot-selective manner (Ye et al., 2015). It seems likely that tamoxifen treatment will perturb the biology of many more cells and tissues with functional estrogen receptors, and more reports of tamoxifen toxicity or unexpected biology can be expected. Alternatives to tamoxifen that are used for regulation of conditional gene expression include members of the tetracycline family, such as doxycycline. However, antibiotic treatment almost always produces transient dysbiosis with potential metabolic implications. Furthermore, even when used at very low concentrations, tetracyclines induce mitochondrial proteotoxic stress and alter mitochondrial function in many tissues and organisms (Moullan et al., 2015), findings that may be very relevant for interpretation of metabolic phenotypes.

Equally pervasive and problematic is the ongoing use of non-littermate controls in studies of transgenic or knockout mice. Failure to use littermate controls introduces tremendous potential for phenotypic variability, producing experimental differences between groups that reflect the different genetic backgrounds under study. Although genetic and phenotypic differences between C57BL/6J and C57BL/6N substrains have been extensively described (Mekada et al., 2009), many scientists remain unaware of the precise genetic background of commercially purchased mice used in their own metabolism studies. Although the use of littermate controls is more expensive and time intensive, it substantially diminishes the likelihood of reporting artifactual differences between control and experimental groups (Sigmund, 2000). These examples likely represent only a small visible portion of the iceberg that can sink many noble hypotheses and papers if the proper controls are not always carefully deployed and analyzed simultaneously in each experiment.

## Mice Are Not Always Good Models for Studying Disease Pathophysiology Relevant to Humans

If substantial challenges limit the universal reproducibility of many preclinical studies, even greater obstacles arise in reproducing and translating key findings from preclinical studies to humans. Much has been written about the poor reproducibility of preclinical studies in oncology (Begley and Ellis, 2012) and cardiovascular biology (Libby, 2015); however, the field of translational metabolism may not be so different. Inflammation is a powerful driver of murine metabolic phenotypes such as islet dysfunction, non-alcoholic fatty liver disease, atherosclerosis, and insulin resistance. The extent to which inflammation similarly drives the pathophysiology of these disorders in humans is more challenging to ascertain. Evaluation of gene expression profiles in multiple distinct mouse models of liver inflammation versus profiling of RNA from human liver biopsies obtained from patients with NASH revealed substantial inter-species differences and surprisingly little overlap in gene expression profiles (Teufel et al., 2016). Similar species-specific differences have become evident in studies of anti-inflammatory interventions for the treatment of diabetes, which often produce robust resolution of experimental insulin resistance and diabetes in many animal models (Donath, 2014). On the other hand, targeting of interleukin-1 or tumor necrosis factor-α or the NF-κB pathway in humans produced relatively modest, often clinically insignificant, improvements in insulin secretion and glucose control (Donath, 2016). Hence, it remains challenging to produce robust and clinically meaningful translation of immune interventions for metabolic disorders from animals to humans.

Recent studies also highlight the limitations inherent in broadly extrapolating data from inbred mice, housed at certain temperatures, to other animals or species. The temperature of the animal facility has profound effects on metabolically sensitive tissues, not just on the extent of beige or brown adipose tissue activation (Speakman and Keijer, 2012); temperature may also modify the development of tissue and systemic inflammation and experimental atherosclerosis (Tian et al., 2016). Furthermore, the immune system of mice raised in pathogen-free barrier facilities is notably different compared to the innate and adaptive immune systems characterized in pet store mice (Beura et al., 2016), findings with clear implications for development of immune interventions that can be translated from the laboratory into the clinic. Tremendous differences in metabolic rate, basal cardiovascular function, feeding behavior, hepatic lipid metabolism, and other species-specific physiological differences may also contribute to difficulties in translation of preclinical research findings across species.

## Translational Challenges in Enteroendocrine Biology

The success in development of gut hormone-based therapies for the treatment of diabetes and obesity has re-energized the field of enteroendocrine cell (EEC) biology. However, the complexity of multiple EEC populations within the small and large bowel, coupled with important species-specific differences in the molecular characterization and function of EECs, raises important caveats for translation and drug development (Drucker, 2016). EEC scientists have long been aware that some regulators of GLP-1 secretion in rodents, such as GIP, gastrin-releasing peptide, leptin, insulin, artificial sweeteners, and other neurotransmitters, do not appear to robustly stimulate GLP-1 secretion in humans (Lim and Brubaker, 2006; Pais et al., 2016).

A number of molecules targeting gut hormone secretion have worked beautifully in rodents, but not so well in humans. Multiple companies (at least five) tested the clinical efficacy of chemically distinct GPR119 agonists in subjects with T2D, following promising activity (stimulation of incretin and insulin secretion) of these same GPR119 agonists in studies of diabetic mice and rats. The clinical results in normal and diabetic human subjects were uniformly disappointing, and no GPR119 agonist progressed beyond phase 2 evaluation in clinical trials of subjects with T2D (Ritter et al., 2016). Although the lead GPR40 agonist, TAK875, was discontinued primarily for reasons related to hepatic toxicity, preclinical findings with GPR40 demonstrated robust stimulation of incretin and insulin secretion; however, not all of these findings were reproduced in human studies (Mancini and Poitout, 2015). The importance of species-specific differences in receptor signaling, differential effects of full versus biased versus allosteric agonism, and differential actions of GPR40 agonists on rodent versus human enteroendocrine cells requires more extensive evaluation (Mancini and Poitout, 2015).

Most academic investigators do not have the resources required to simultaneously assess the efficacy of promising therapeutics in both preclinical and clinical studies. Hodge and colleagues developed a therapeutic mixture of four different natural compounds targeting gut-related mechanisms linked to weight loss and glucoregulation. Each individual compound exhibited favorable pharmacological activity, including the stimulation of EEC and GLP-1 secretion, and activation of fatty acid receptors, in preclinical studies (Hodge et al., 2016). GSK457 was developed as a mixture of these four individual compounds, oligofructosaccharide, apple pectin, blackcurrant extract, and oleic acid, combined in a ratio of 5:5:2:3. GSK457, when administered alone or in combination with a GLP-1R agonist, produced robust weight loss, reductions in glycemia, and amelioration of hepatic steatosis in high-fat diet-fed or db/db mice. In contrast, the same investigators assessed the efficacy of GSK457 in three different groups of human subjects with or without T2D, on top of baseline metformin or liraglutide therapy, for 6 weeks. Disappointingly, no meaningful improvement in body weight or glycemic parameters was detected in subjects treated with GSK457 (Hodge et al., 2016), highlighting further challenges in translating robust results from preclinical studies into the clinic.

## Young versus Older Animals and Likelihood of Translation

Even after avoiding many of the pitfalls enumerated above, the most carefully done science in rodents may fail to translate in other species or humans, due to considerable inherent physiological differences in the biology of small versus larger animals and rodents versus humans. For obvious reasons of availability, cost, and efficiency (time to complete an experiment), young mice, often 2–6 months of age, are widely used in studies of preclinical research. Although T2D has sadly become more prevalent in our children, the majority of human subjects with T2D are usually much older, often in the fifth through ninth decade of life. As a result, many older human subjects have experienced years of low-grade tissue inflammation and fibrosis, dyslipidemia, weight gain, and hypertension, associated with a gradual progression from impaired glucose tolerance to frank dysglycemia and T2D. The suitability of using young mice, often predominantly only one strain (C57BL/6J), for assessing the translational potential of new therapeutic mechanisms is questionable. Younger animals are far more likely to exhibit a greater potential for organ repair, cellular plasticity, and cell proliferation, compared to older animals. Indeed the field of neuroscience research is replete with extensive reports of age-associated reductions in cognition and synaptic plasticity, and it is generally easier to prevent disease in young mice (T1D in NOD mice, atherosclerosis in *apoE*$^{-/-}$ mice) than reverse established disease in much older animals. Nevertheless, there is no consistent expectation from funding agencies or journals that key novel and compelling results, frequently touted as having exciting translational potential, be examined critically not only in younger mice, but also in older animals with established disease.

## The Promise and Challenge of Therapeutically Targeting Islet Cells

The incretin field contains hundreds of papers, including several from our own laboratory, describing robust expansion of β cell mass in diabetic mice and rats treated with GLP-1R agonists and DPP-4 inhibitors. Indeed, a single injection of GLP-1 stimulates β cell proliferation in young mice and rats, and GLP-1R signaling also inhibits β cell death not only in rodent, but also in human islets (Campbell and Drucker, 2013). These findings raised considerable expectations in the clinical diabetes world surrounding the potential for disease modification, based on the hope that incretin-based therapy might similarly expand or preserve functional β cell mass in human subjects with T2D.

After more than a decade of clinical trials with GLP-1R agonists and DPP-4 inhibitors, the promise of disease modification and preferential recovery or preservation of β cell function in human subjects with T2D remains unfulfilled (Campbell and Drucker, 2013; Drucker, 2011). We now understand that at least some of the translational gap reflects striking age-associated differences in the capacity for β cell proliferation in rodents. Older β cells (from rats or mice greater than 6 months of age) exhibit limited replicative potential, including lack of proliferative responses to GLP-1R agonists (Rankin and Kushner, 2009; Tschen et al., 2009). Moreover, unlike promising findings with islets isolated from young rodents, it remains very challenging to demonstrate that human β cells exhibit a robust proliferative response to GLP-1R agonists under most experimental conditions ex vivo (Parnaud et al., 2008).

Notwithstanding the behavior of older islets, many studies using human islets employ small numbers of islets from a limited number of donors, reflecting the challenges most laboratories encounter in human islet experimentation. However, even the most experienced laboratories and consortiums report considerable heterogeneity in human islet insulin content and secretory capacity, emphasizing the importance of sufficiently powered studies for generation of robust and reproducible human islet data (Kayton et al., 2015; Lyon et al., 2016). The recognition that human islets also exhibit profound age-associated profiles of gene expression linked to differences in replicative capacity (Arda et al., 2016) provides additional guidance for investigators wishing to study the translational biology of older diabetic human islets.

The proliferative, anti-inflammatory, and anti-apoptotic properties of incretin-based therapies in rodent islets have also been observed in studies using NOD mice, where multiple laboratories have documented prevention of disease development or improvement of murine or human islet graft survival using GLP-1R agonists or DPP-4 inhibitors. On the other hand, we were unable to demonstrate any meaningful attenuation of diabetes development using exendin-4 in NOD mice (Hadjiyanni et al., 2008), and efforts to date to demonstrate the clinical utility of GLP-1R agonists in subjects with T1D, with and without immunosuppression, have not been successful (Rother et al., 2009). Similarly, anecdotal reports using GLP-1R agonists in small numbers of subjects with T1D post islet transplant abound; however, no large randomized controlled trials have demonstrated beneficial effects of incretin-based therapies in the islet transplant setting. Although the NOD mouse is widely regarded as the best available preclinical model for the study of T1D, whether the pathophysiology of insulitis and immune-mediated β cell destruction elegantly described in the NOD mouse is an excellent model for translational intervention studies in T1D continues to be debated (Battaglia and Atkinson, 2015). Nevertheless, despite a series of disappointments, efforts to demonstrate the potential utility of incretin-based and other regenerative and immunotherapies as components of a multi-pronged approach toward preservation of β cell function continue, and one hopes that past experimental disappointments do not preclude more promising results in future studies.

## Journals, Editors, Public Relations Staff, and the Media

A few years ago in 2009, several years after the clinical introduction of incretin-based therapies, I was startled to read a press release from a prestigious medical center containing recommendations to restrict the clinical use of incretin-based therapies, based on findings in a very small number of rats in a preclinical study. The controversy surrounding the expression and biological activity of the GLP-1R in normal and neoplastic rodent and human pancreatic tissue continued for several years and was also associated with inappropriate statistical analysis, questionable interpretation of data from adverse event reporting system databases, inadequately controlled animal studies, and technically flawed analysis of a small number of human histology samples. Nevertheless, many of these stories with inappropriate clinical conclusions were published in respected journals and widely disseminated in the media, with feature stories in the New York Times and other leading media outlets. Collectively, a series of peer-reviewed papers contained a substantial amount of alarmist misinformation, coupled with recommendations often trumpeted by accompanying editorials that incretin-based therapy should be curtailed or withdrawn. These publications and their conclusions supported the filing of lawsuits and led to much patient anxiety, ultimately requiring the attention of regulatory authorities (Egan et al., 2014). After a huge amount of time and financial resources expended in independent attempts to reproduce many of the key findings, regulatory agencies concluded that many of the original scientific reports alleging serious safety issues were suboptimal and key conclusions could not be independently reproduced (Bonner-Weir et al., 2014; Drucker, 2013; Egan et al., 2014). Long-term outcome studies in human subjects have subsequently validated the safety and therapeutic benefit of several incretin-based therapies (Drucker, 2016).

The rising tide of angst and discussions related to the fidelity and reproducibility of preclinical research is likely to foster greater awareness of methodological pitfalls and challenges in experimental design and execution. A growing chorus of scientific societies, independent funding agencies and journals, and independent investigators have taken up the rallying cry, and thoughtful recommendations and consensus statements continue to accumulate. These actions alone are unlikely to make a major dent in the status quo, which continues to be problematic (Macleod et al., 2015). Behavioral economics teaches us that a mixture of incentives and penalties stand the best chance of producing meaningful change in the way we do research.

Although scientists must take ultimate responsibility for their standards, ethics, training environments, validation of reagents, experimental results, and reproducibility of their published data, the complex web of individuals contributing to the current reproducibility crisis is worth mentioning. Journal editors attend many meetings, socialize and network with hundreds of scientists, and compete for the most exciting papers from the best labs. Editors in turn are aware that the most tantalizing papers will garner the greatest media visibility and, ultimately, will indirectly attract more page views, advertising revenue, and boost the journal's reputation, impact factor, and profitability. Corresponding authors will not infrequently receive a small editorial nudge if they have not sufficiently framed the biomedical translational importance of their basic science with sufficient positivity, panache, and verve.

The same scientist will, at many institutions, receive regular email requests, periodic visits, and exhortations from public relations or media communications officers, anxious for new exciting stories to tell that highlight the wonderful work being done within the institution. There are monthly newsletters to fill, websites to update, fundraising pitches and portfolios to embellish, and local media contacts always need a new story. 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.

How did we arrive at this state of affairs? Although it is admittedly more difficult to become an elite fighter pilot, astronaut, or head of state, competition for faculty positions and resources in the best academic institutions is fierce, and the most valuable currency continues to be a mixture of publications in ‘‘the best journals,’’ ideally coupled with already secured independent funding. To obtain these valuable prestigious publications, one must meet the standards and expectations of journal editors, who similarly prize research that is spectacular, highly novel, and ideally accompanied by well-defined reductionist mechanisms and immediate obvious translational relevance. Given these challenges, it is perhaps not surprising that most research in my own area of gut hormone action has not been published in the top journals. Perhaps this reflects a degree of mediocrity and lack of scientific talent and imagination (mea culpa) in most individuals who have pursued the biology of regulatory peptides. Alternatively, it may reflect a historical tradition in the field of careful incremental physiology, studying how peptides that circulate at very low levels engage intricate communication mechanisms, in part through neuronal pathways, which may be challenging to tease out and simplify. Nevertheless, my plodding colleagues and I have witnessed the development, approval, and clinical utilization of several new drug classes for

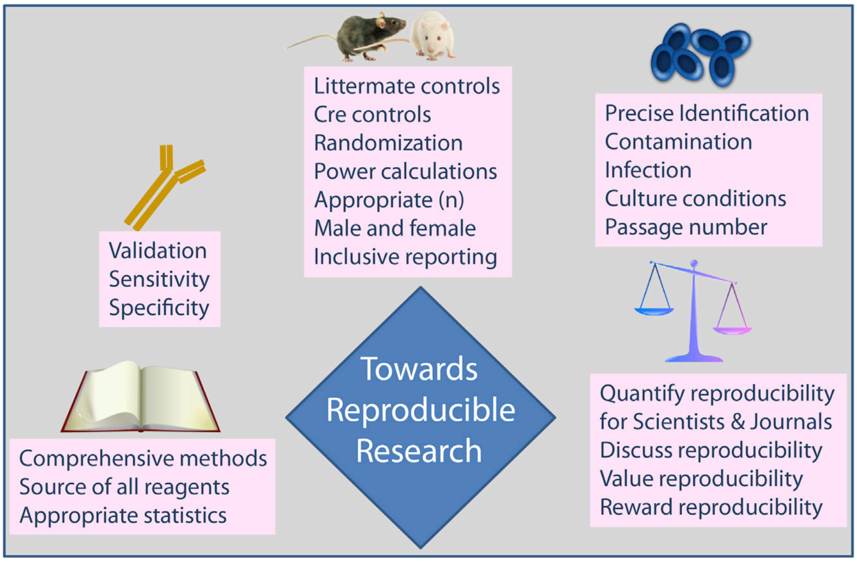

**Figure 2. Issues Contributing to Suboptimal Reproducibility of Preclinical Research Are Highlighted**
Strategies to enhance research reproducibility are outlined.

diabetes, obesity, and gastrointestinal disease, working in a field where 99% of the papers are published in "mid-tier," yet respectable, physiology, biochemistry, endocrinology, and gastroenterology journals. Hence, despite a paucity of high-impact papers in the best journals, it seems clear that careful incremental, solid science, although rarely flashy, may, brick by brick, help build a field of science that is reproducible within and across many species, ultimately enabling successful drug development programs (Drucker, 2015).

## Moving beyond the Status Quo toward Highly Reproducible Research

While accepting the notion that sunlight is the best disinfectant for many problems, simply highlighting existing challenges (Figure 2) in an anecdotal way, or publishing guidelines, commentaries, or position papers, while perhaps helpful, seems unlikely to move the needle in a meaningful way. Below, I provide some simple suggestions, including recommendations made by others that may be useful for enhancing research reproducibility (Figure 2).

### *Transparency in Communication*

A very simple strategy for enabling research to be more easily reproduced is the provision of sufficient experimental detail, including careful description of and source of reagents, cell lines, and animals used in each experiment. While this sounds obvious, assessment of the extent to which publications routinely provide sufficient information to facilitate reproducibility reveals major gaps in our collective communication styles. Vasilevsky and colleagues surveyed 238 journal articles across five biomedical research disciplines and found that 54% of the research resources used to carry out experiments were not adequately described or identified (Vasilevsky et al., 2013). Simply enhancing journal requirements for more detailed research reagent identification and reporting would increase the likelihood that future scientists were using the same reagents, conditions, and animals, in attempts to reproduce key findings. Although some journals cite space limitations, precluding provision of more extensive information, the widespread use of online supplemental information should facilitate, not hinder, precise research communication.

### *Feature Discussions of Research Reproducibility at Scientific Meetings*

We all attend scientific meetings to hear and present the most updated science, network, and learn new techniques and best practices from colleagues. Although sessions are often devoted to challenges and pitfalls inherent to specific research areas, it is rare to see sessions dedicated to reproducibility issues in research. Imagine if most scientific meetings held a regular panel discussion with representation from scientists, funding agencies, and journal editors devoted to their recent experiences with fostering research reproducibility. Common problems could be highlighted and solutions proposed. Since new methodological issues and techniques surface constantly, the annual "reproducibility" symposia need not become stale or repetitive. Simply ingraining the importance of regularly discussing reproducibility problems and identifying methods to improve our research practices would make us more accountable to each other within our own scientific ecosystems.

### *Design and Reporting of Animal Experiments*

Extensive recommendations have been developed for design and reporting of animal experiments in the form of the "Animal Research: Reporting of in Vivo Experiments" (ARRIVE) Guidelines. These recommendations are comprehensive and span study design, housing, timing of experiments, more extensive description and reporting of animals used, animal husbandry, sample size calculations, randomization and analysis and reporting of outcomes, including all positive and negative results, adverse events, and use of appropriate statistical methods (Kilkenny et al., 2010). Despite widespread endorsement of the guidelines by funding agencies and multiple journals, the impact to date of the guidelines on improvement of reporting of animal experiments has been modest (Baker et al., 2014). Although some journals have developed checklists of required information that must accompany each article, most of these checklists fall short in regard to reporting requirements recommended for preclinical studies.

It is noteworthy that in the clinical trial domain, most journals, review boards, and funding agencies now enforce mandatory registration of clinical trials, with obligatory reporting requirements, on Clinicaltrials.gov. This allows for assessment of the design, pre-specified outcomes, and, ideally, results of a clinical trial, with information on results sometimes provided in advance of a peer-reviewed publication. In contrast, the details surrounding ongoing animal experimentation within laboratories and institutions are opaque and not accessible. It is not uncommon for scientists testing the efficacy of a new therapeutic agent to try numerous doses and concentrations and time points, employing different modes of administration in multiple animal models, using mice of different ages and health status, to finally land on the experiment that "works the best," and this one experiment makes it into the paper as a key figure. None of the dozens of experiments that did not work are ever divulged.

It is expected and understandable that each envisioned putative experimental result (often a therapeutic response) will require exploration and refinement of pharmacokinetic and pharmacodynamic relationships unique to each reagent. Nevertheless, it also seems likely that in some instances, the experiment simply does not work most of the time, the hypothesis is generally not sustained, and the scientist simply keeps on searching for just the right conditions that will provide the desired results, however narrowly applicable they might be. Reporting of negative results (conditions and models tried where an expected result was not obtained) is likely to be extremely valuable to the research community, yet at present there is no uniformly accepted or promoted mechanism enabling or requiring reporting of negative results. These unpublished experiments are not restricted to the academic sector, as the pharmaceutical industry also undertakes a great deal of preclinical research that may never be publicly disclosed. Scientists would likely be disinclined to volunteer such information, unless mandated to do so by funding agencies, institutional guidelines, animal care committees, or journals. After all, who wants to disclose that one's exciting new therapeutic agent or mechanism only worked in 10% of the experiments it was tested in? Nevertheless, for scientists, journals, and funding agencies serious about reproducibility, a mechanism or channel for reporting all results, including negative results, would go a long way toward enhancing reproducibility and likely save colleagues a great deal of time and financial resources. The increasing recognition that questioning existing results or reporting negative results has great value has fostered innovation in discussions of reproducibility through websites such as PubPeer (https://pubpeer.com/) or the Preclinical Reproducibility and Robustness publication channel developed by F1000Research (http://f1000research.com/channels/PRR). While these online forums are in their infancy, have growing pains, and may foster problematic discussions and accusations, the quality of the comments and motives of the participants should become more refined and improve over time. In the future, it seems likely that journals will develop their own portals, linked to individual papers, to facilitate the updating of published results and the reporting and tracking of efforts directed at reproducing key published findings.

## Tracking Reproducibility; the R Index

The McNamara or quantitative fallacy states that decisions should only be made based on quantitative data, because if you can't measure something it may not exist. Similarly, Michael Bloomberg ran his business and New York City in part by reminding colleagues "In God we trust. Everyone else bring data."

A combination of transparency and accountability for one's track record is an accepted part of how we judge scientists; however, measurement of reproducibility is currently missing from most metrics and equations used to evaluate scientists. Although all scientists are data driven, we currently have no accepted way to track and quantify our own track record of scientific reproducibility. This reproducibility knowledge gap extends to our funding agencies and journals, which collectively have little understanding of whether the science they regularly fund and publish, respectively, turns out to be reproducible. At present, reproducibility is the subject of informal water cooler or bar room discussions, editorial musings, and consensus conferences (McNutt, 2014), but there is little attempt to construct and validate an index for research reproducibility. Simple steps could be taken by granting agencies, who might ask for a one-half page description of the scientists' major reproducible research findings. This requirement would not be applicable to junior scientists, and might kick in after 10 years of independent productivity.

Rather than only quantifying citations, papers in top journals, and research funding, why not quantify reproducibility? Imagine if each scientist was associated with a $R_s$ (Reproducibility$_{Scientist}$) index, reflecting the number of times the key scientific findings in a paper had been reproduced by at least one other independent research group. Of course, one would have to carefully debate and refine the meaning of "reproducible" (Goodman et al., 2016), but perhaps one could start simply by requiring that (a) the key findings and (b) at least 50% of the experimental data, from a single paper, were independently reproduced by at least one other research group. So a senior scientist with a $R_s$ index of 40 would have published 40 research papers with findings found to be independently reproduced by others. Adjudication of the reproducibility of each paper would have to be carefully tracked over time, ideally by an independent body, which would require funding to sustain its activities. This type of undertaking presents major feasibility challenges, but also opportunities for independent organizations dedicated to the reproducibility of published research.

The reproducibility index should not be restricted to scientists. Each journal should also have an associated $R_J$ (Reproducibility$_{Journal}$) index, similarly reflecting the number of papers it publishes that are ultimately found to be reproducible. The R indexes could also be divided by the total number of publications (per scientist or journal) to yield $R_{\%s}$ and $R_{\%J}$ indices, reflecting the proportion of total papers and output ultimately found to be reproducible. Although it would take some time to generate metrics and establish the validity and utility of these measures, what scientist or journal would aspire to have a low reproducibility index? Hiring, promotion, funding, and award deliberations would ideally incorporate assessment of reproducibility as one additional factor to be considered in ranking of candidates. Who would want to fund, hire, or reward a scientist with a low reproducibility index? A potential advantage of this index is that it does not matter whether one regularly publishes only in high impact journals or simply does careful meaningful science published in subspecialty or mid-tier general science journals. Although no metric is likely to be without flaws or critics, pursuing high-quality reproducible science, without formally measuring reproducibility, is not likely to be successful. As Begley and Ioannidis have noted, "We get what we incentivize" (Begley and Ioannidis, 2015), and if we fail to measure and incentivize careful reproducible science, it is unlikely we will change the landscape of our current problematic scientific enterprise. The importance and attractiveness of reproducibility research could be enhanced by having funding agencies allocate a dedicated proportion of new funding for grants directed at research reproducibility, focused on the most potentially transformative findings within a field. Alternatively, companies in the private sector might pool resources toward establishment of small independent reproducibility research laboratories, tasked with reproducing major key findings within a select number of scientific disciplines, with results profiled annually at conferences and in journals. As the editorial leadership of this journal has noted, ensuring reproducibility and experimental replication is the responsibility of the entire scientific community (Emambokus and Granger, 2015), not just someone else's problem.

## "Great Expectations" for Reproducible Scientific Research

In closing, as scientists we are privileged to work in an era blessed with boundless opportunity and constant technological innovation. Our funding agencies and institutions have high expectations, and our fellow citizens willingly divert tens of billions of tax and charitable dollars to support our research enterprise. Do we really want to continue pursuing scientific investigation using methods associated with a high degree of scientific irreproducibility, much of which can be prevented by instituting better scientific practice? Perhaps we should institute a culture of reproducibility within our own laboratories, routinely assigning new lab members a project directed at reproducing a subset of key findings made by other lab members within the past 12 months. Let us take some inspiration from Pip, a Charles Dickens character in *Great Expectations* who felt guilty about not coming clean and proclaimed "In a word, I was too cowardly to do what I knew to be right, as I had been too cowardly to avoid doing what I knew to be wrong." The path to greater reproducibility in preclinical research is one well worth traveling, and as a community of scientists, we should waste no time in embarking on this journey together.

## ACKNOWLEDGMENTS

D.J.D. is supported in part by operating grants from CIHR, a Canada Research Chair in Regulatory Peptides, and a Banting and Best Diabetes Centre-Novo Nordisk Chair in Incretin Biology. D.J.D. has served as an advisor or consultant to Arisaph Pharmaceuticals, Intarcia, Merck Research Laboratories, Novo Nordisk, and Receptos Inc. Mt. Sinai Hospital receives funding from GSK, Merck, Novo Nordisk, and Sanofi Inc. for preclinical studies in the Drucker laboratory.

## REFERENCES

Arda, H.E., Li, L., Tsai, J., Torre, E.A., Rosli, Y., Peiris, H., Spitale, R.C., Dai, C., Gu, X., Qu, K., et al. (2016). Age-Dependent Pancreatic Gene Regulation Reveals Mechanisms Governing Human β Cell Function. Cell Metab. *23*, 909–920.

Ayala, J.E., Bracy, D.P., James, F.D., Burmeister, M.A., Wasserman, D.H., and Drucker, D.J. (2010). Glucagon-like peptide-1 receptor knockout mice are protected from high-fat diet-induced insulin resistance. Endocrinology *151*, 4678–4687.

Baker, M. (2015). Reproducibility crisis: Blame it on the antibodies. Nature *521*, 274–276.

Baker, D., Lidster, K., Sottomayor, A., and Amor, S. (2014). Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. PLoS Biol. *12*, e1001756.

Battaglia, M., and Atkinson, M.A. (2015). The streetlight effect in type 1 diabetes. Diabetes *64*, 1081–1090.

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

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

Bersell, K., Choudhury, S., Mollova, M., Polizzotti, B.D., Ganapathy, B., Walsh, S., Wadugu, B., Arab, S., and Kühn, B. (2013). Moderate and high amounts of tamoxifen in αMHC-MerCreMer mice induce a DNA damage response, leading to heart failure and death. Dis. Model. Mech. *6*, 1459–1469.

Beura, L.K., Hamilton, S.E., Bi, K., Schenkel, J.M., Odumade, O.A., Casey, K.A., Thompson, E.A., Fraser, K.A., Rosato, P.C., Filali-Mouhim, A., et al. (2016). Normalizing the environment recapitulates adult human immune traits in laboratory mice. Nature *532*, 512–516.

Bonner-Weir, S., In't Veld, P.A., and Weir, G.C. (2014). Reanalysis of study of pancreatic effects of incretin therapy: methodological deficiencies. Diabetes Obes. Metab. *16*, 661–666.

Bradbury, A., and Plückthun, A. (2015). Reproducibility: Standardize antibodies used in research. Nature *518*, 27–29.

Brouwers, B., de Faudeur, G., Osipovich, A.B., Goyvaerts, L., Lemaire, K., Boesmans, L., Cauwelier, E.J., Granvik, M., Pruniau, V.P., Van Lommel, L., et al. (2014). Impaired islet function in commonly used transgenic mouse lines due to human growth hormone minigene expression. Cell Metab. *20*, 979–990.

Campbell, J.E., and Drucker, D.J. (2013). Pharmacology, physiology, and mechanisms of incretin hormone action. Cell Metab. *17*, 819–837.

Claus, S.P., Tsang, T.M., Wang, Y., Cloarec, O., Skordi, E., Martin, F.P., Rezzi, S., Ross, A., Kochhar, S., Holmes, E., and Nicholson, J.K. (2008). Systemic multicompartmental effects of the gut microbiome on mouse metabolic phenotypes. Mol. Syst. Biol. *4*, 219.

Clayton, J.A., and Collins, F.S. (2014). Policy: NIH to balance sex in cell and animal studies. Nature *509*, 282–283.

Collins, F.S., and Tabak, L.A. (2014). Policy: NIH plans to enhance reproducibility. Nature *505*, 612–613.

Donath, M.Y. (2014). Targeting inflammation in the treatment of type 2 diabetes: time to start. Nat. Rev. Drug Discov. *13*, 465–476.

Donath, M.Y. (2016). Multiple benefits of targeting inflammation in the treatment of type 2 diabetes. Diabetologia *59*, 679–682.

Drucker, D.J. (2011). Incretin-based therapy and the quest for sustained improvements in β-cell health. Diabetes Care *34*, 2133–2135.

Drucker, D.J. (2013). Incretin action in the pancreas: potential promise, possible perils, and pathological pitfalls. Diabetes *62*, 3316–3323.

Drucker, D.J. (2015). Deciphering metabolic messages from the gut drives therapeutic innovation: the 2014 Banting Lecture. Diabetes *64*, 317–326.

Drucker, D.J. (2016). Evolving Concepts and Translational Relevance of Enteroendocrine Cell Biology. J. Clin. Endocrinol. Metab. *101*, 778–786.

Drucker, D.J., and Yusta, B. (2014). Physiology and pharmacology of the enteroendocrine hormone glucagon-like peptide-2. Annu. Rev. Physiol. *76*, 561–583.

Egan, A.G., Blind, E., Dunder, K., de Graeff, P.A., Hummer, B.T., Bourcier, T., and Rosebraugh, C. (2014). Pancreatic safety of incretin-based drugs–FDA and EMA assessment. N. Engl. J. Med. *370*, 794–797.

Emambokus, N., and Granger, A. (2015). The Elephant in the Room. Cell Metab. *22*, 345.

Freedman, L.P., Gibson, M.C., Ethier, S.P., Soule, H.R., Neve, R.M., and Reid, Y.A. (2015). Reproducibility: changing the policies and culture of cell line authentication. Nat. Methods *12*, 493–497.

Freedman, L.P., Gibson, M.C., Bradbury, A.R.M., Buchberg, A.M., Davis, D., Dolled-Filhart, M.P., Lund-Johansen, F., and Rimm, D.L. (2016). [Letter to the Editor] The need for improved education and training in research antibody usage and validation practices. Biotechniques *61*, 16–18.

Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature *536*, 41–47.

Goodman, S.N., Fanelli, D., and Ioannidis, J.P. (2016). What does research reproducibility mean? Sci. Transl. Med. *8*, 341ps12.

Gore, A.C. (2013). Editorial: antibody validation requirements for articles published in endocrinology. Endocrinology *154*, 579–580.

Hadjiyanni, I., Baggio, L.L., Poussier, P., and Drucker, D.J. (2008). Exendin-4 modulates diabetes onset in nonobese diabetic mice. Endocrinology *149*, 1338–1349.

Hansotia, T., Maida, A., Flock, G., Yamada, Y., Tsukiyama, K., Seino, Y., and Drucker, D.J. (2007). Extrapancreatic incretin receptors modulate glucose homeostasis, body weight, and energy expenditure. J. Clin. Invest. *117*, 143–152.

Harwood, H.J., Jr., Listrani, P., and Wagner, J.D. (2012). Nonhuman primates and other animal models in diabetes research. J. Diabetes Sci. Technol. *6*, 503–514.

Heffner, C.S., Herbert Pratt, C., Babiuk, R.P., Sharma, Y., Rockwood, S.F., Donahue, L.R., Eppig, J.T., and Murray, S.A. (2012). Supporting conditional mouse mutagenesis with a comprehensive cre characterization resource. Nat. Commun. *3*, 1218.

Hodge, R.J., Paulik, M.A., Walker, A., Boucheron, J.A., McMullen, S.L., Gillmor, D.S., and Nunez, D.J. (2016). Weight and Glucose Reduction Observed with a Combination of Nutritional Agents in Rodent Models Does Not Translate to Humans in a Randomized Clinical Trial with Healthy Volunteers and Subjects with Type 2 Diabetes. PLoS ONE *11*, e0153151.

Hrabě de Angelis, M., Nicholson, G., Selloum, M., White, J.K., Morgan, H., Ramirez-Solis, R., Sorg, T., Wells, S., Fuchs, H., Fray, M., et al.; EUMODIC Consortium (2015). Analysis of mammalian gene function through broad-based phenotypic screens across a consortium of mouse clinics. Nat. Genet. *47*, 969–978.

Huh, W.J., Khurana, S.S., Geahlen, J.H., Kohli, K., Waller, R.A., and Mills, J.C. (2012). Tamoxifen induces rapid, reversible atrophy, and metaplasia in mouse stomach. Gastroenterology *142*, 21–24.e7.

Kayton, N.S., Poffenberger, G., Henske, J., Dai, C., Thompson, C., Aramandla, R., Shostak, A., Nicholson, W., Brissova, M., Bush, W.S., and Powers, A.C. (2015). Human islet preparations distributed for research exhibit a variety of insulin-secretory profiles. Am. J. Physiol. Endocrinol. Metab. *308*, E592–E602.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F., Cuthill, I.C., Fry, D., Hutton, J., and Altman, D.G. (2009). Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoS ONE *4*, e7824.

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., and Altman, D.G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. *8*, e1000412.

Koitabashi, N., Bedja, D., Zaiman, A.L., Pinto, Y.M., Zhang, M., Gabrielson, K.L., Takimoto, E., and Kass, D.A. (2009). Avoidance of transient cardiomyopathy in cardiomyocyte-targeted tamoxifen-induced MerCreMer gene deletion models. Circ. Res. *105*, 12–15.

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. Nature *490*, 187–191.

Lee, J.Y., Ristow, M., Lin, X., White, M.F., Magnuson, M.A., and Hennighausen, L. (2006). RIP-Cre revisited, evidence for impairments of pancreatic beta-cell function. J. Biol. Chem. *281*, 2649–2653.

Libby, P. (2015). Murine “model” monotheism: an iconoclast at the altar of mouse. Circ. Res. *117*, 921–925.

Lim, G.E., and Brubaker, P.L. (2006). Glucagon-Like Peptide 1 Secretion by the L-Cell: The View From Within. Diabetes *55*, S70–S77.

Lyon, J., Manning Fox, J.E., Spigelman, A.F., Kim, R., Smith, N., O’Gorman, D., Kin, T., Shapiro, A.M., Rajotte, R.V., and MacDonald, P.E. (2016). Research-Focused Isolation of Human Islets From Donors With and Without Diabetes at the Alberta Diabetes Institute IsletCore. Endocrinology *157*, 560–569.

Macleod, M.R., Lawson McLean, A., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., Hirst, T., Hemblade, R., Bahor, Z., Nunes-Fonseca, C., et al. (2015). Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol. *13*, e1002273.

Magnuson, M.A., and Osipovich, A.B. (2013). Pancreas-specific Cre driver lines and considerations for their prudent use. Cell Metab. *18*, 9–20.

Mancini, A.D., and Poitout, V. (2015). GPR40 agonists for the treatment of type 2 diabetes: life after ‘TAKing’ a hit. Diabetes Obes. Metab. *17*, 622–629.

McNutt, M. (2014). Journals unite for reproducibility. Science *346*, 679.

Mekada, K., Abe, K., Murakami, A., Nakamura, S., Nakata, H., Moriwaki, K., Obata, Y., and Yoshiki, A. (2009). Genetic differences among C57BL/6 substrains. Exp. Anim. *58*, 141–149.

Moullan, N., Mouchiroud, L., Wang, X., Ryu, D., Williams, E.G., Mottis, A., Jovaisaite, V., Frochaux, M.V., Quiros, P.M., Deplancke, B., et al. (2015). Tetracyclines Disturb Mitochondrial Function across Eukaryotic Models: A Call for Caution in Biomedical Research. Cell Rep., S2211-1247(15)00180-1.

Pais, R., Gribble, F.M., and Reimann, F. (2016). Stimulation of incretin secreting cells. Ther. Adv. Endocrinol. Metab. *7*, 24–42.

Panjwani, N., Mulvihill, E.E., Longuet, C., Yusta, B., Campbell, J.E., Brown, T.J., Streutker, C., Holland, D., Cao, X., Baggio, L.L., and Drucker, D.J. (2013). GLP-1 receptor activation indirectly reduces hepatic lipid accumulation but does not attenuate development of atherosclerosis in diabetic male ApoE(-/-) mice. Endocrinology *154*, 127–139.

Parnaud, G., Bosco, D., Berney, T., Pattou, F., Kerr-Conte, J., Donath, M.Y., Bruun, C., Mandrup-Poulsen, T., Billestrup, N., and Halban, P.A. (2008). Proliferation of sorted human and rat beta cells. Diabetologia *51*, 91–100.

Prinz, F., Schlange, T., and 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.

Pyke, C., and Knudsen, L.B. (2013). The glucagon-like peptide-1 receptor–or not? Endocrinology *154*, 4–8.

Pyke, C., Heller, R.S., Kirk, R.K., Ørskov, C., Reedtz-Runge, S., Kaastrup, P., Hvelplund, A., Bardram, L., Calatayud, D., and Knudsen, L.B. (2014). GLP-1 receptor localization in monkey and human tissue: novel distribution revealed with extensively validated monoclonal antibody. Endocrinology *155*, 1280–1290.

Rankin, M.M., and Kushner, J.A. (2009). Adaptive beta-cell proliferation is severely restricted with advanced age. Diabetes *58*, 1365–1372.

Ritter, K., Buning, C., Halland, N., Pöverlein, C., and Schwink, L. (2016). G Protein-Coupled Receptor 119 (GPR119) Agonists for the Treatment of Diabetes: Recent Progress and Prevailing Challenges. J. Med. Chem. *59*, 3579–3592.

Rother, K.I., Spain, L.M., Wesley, R.A., Digon, B.J., 3rd, Baron, A., Chen, K., Nelson, P., Dosch, H.M., Palmer, J.P., Brooks-Worrell, B., et al. (2009). Effects of exenatide alone and in combination with daclizumab on beta-cell function in long-standing type 1 diabetes. Diabetes Care *32*, 2251–2257.

Schmidt-Supprian, M., and Rajewsky, K. (2007). Vagaries of conditional gene targeting. Nat. Immunol. *8*, 665–668.

Scrocchi, L.A., and Drucker, D.J. (1998). Effects of aging and a high fat diet on body weight and glucose control in glucagon-like peptide-1 receptor -/- mice. Endocrinology *139*, 3127–3132.

Sigmund, C.D. (2000). Viewpoint: are studies in genetically altered mice out of control? Arterioscler. Thromb. Vasc. Biol. *20*, 1425–1429.

Speakman, J.R., and Keijer, J. (2012). Not so hot: Optimal housing temperatures for mice to mimic the thermal environment of humans. Mol. Metab. *2*, 5–9.

Teufel, A., Itzel, T., Erhart, W., Brosch, M., Wang, X.Y., Kim, Y.O., von Schönfels, W., Herrmann, A., Brückner, S., Stickel, F., et al. (2016). Comparison of Gene Expression Patterns Between Mouse Models of Nonalcoholic Fatty Liver Disease and Liver Tissues From Patients. Gastroenterology. S0016-5085(16)34622-4. http://dx.doi.org/10.1053/j.gastro.2016.05.051.

Tian, X.Y., Ganeshan, K., Hong, C., Nguyen, K.D., Qiu, Y., Kim, J., Tangirala, R.K., Tontonoz, P., and Chawla, A. (2016). Thermoneutral Housing Accelerates Metabolic Inflammation to Potentiate Atherosclerosis but Not Insulin Resistance. Cell Metab. *23*, 165–178.

Tschen, S.I., Dhawan, S., Gurlo, T., and Bhushan, A. (2009). Age-dependent decline in beta-cell proliferation restricts the capacity of beta-cell regeneration in mice. Diabetes *58*, 1312–1320.

Ussar, S., Griffin, N.W., Bezy, O., Fujisaka, S., Vienberg, S., Softic, S., Deng, L., Bry, L., Gordon, J.I., and Kahn, C.R. (2015). Interactions between Gut Microbiota, Host Genetics and Diet Modulate the Predisposition to Obesity and Metabolic Syndrome. Cell Metab. *22*, 516–530.

Vasandani, C., Clark, G., Adams-Huet, B., Quittner, C., and Garg, A. (2016). Efficacy of Metreleptin in Patients with Type 1 Diabetes. In 2016 Annual

Meeting of the American Diabetes Association (New Orleans).

Vasilevsky, N.A., Brush, M.H., Paddock, H., Ponting, L., Tripathy, S.J., Larocca, G.M., and Haendel, M.A. (2013). On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ *1*, e148.

Wang, B., Chandrasekera, P.C., and Pippin, J.J. (2014). Leptin- and leptin receptor-deficient rodent models: relevance for human type 2 diabetes. Curr. Diabetes Rev. *10*, 131–145.

Wicksteed, B., Brissova, M., Yan, W., Opland, D.M., Plank, J.L., Reinert, R.B., Dickson, L.M., Tamarina, N.A., Philipson, L.H., Shostak, A., et al. (2010). Conditional gene targeting in mouse pancreatic ß-Cells: analysis of ectopic Cre transgene expression in the brain. Diabetes *59*, 3090–3098.

Woods, S.C., and Begg, D.P. (2015). Food for Thought: Revisiting the Complexity of Food Intake. Cell Metab. *22*, 348–351.

Ye, R., Wang, Q.A., Tao, C., Vishvanath, L., Shao, M., McDonald, J.G., Gupta, R.K., and Scherer, P.E. (2015). Impact of tamoxifen on adipocyte lineage tracing: Inducer of adipogenesis and prolonged nuclear translocation of Cre recombinase. Mol. Metab. *4*, 771–778.