FORUM

Raising the Bar for Reproducible Science at the U.S. Environmental Protection Agency Office of Research and Development

Barbara Jane George*, Jon R. Sobus[†], Lara P. Phelps[‡], Brenda Rashleigh[§], Jane Ellen Simmons*, Ronald N. Hines[*,1], and the Community of Practice for Statistics Guidance Documents Working Groups

*US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory; [†]US Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory; [‡]US Environmental Protection Agency, Office of Research and Development, Office of the Science Advisor, Research Triangle Park, North Carolina 27711; and [§]US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Narragansett, Rhode Island 02882

[1]To whom correspondence should be addressed at US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, 109 TW Alexander Dr., MD 305-02, Research Triangle Park, NC 27711, E-mail: hines.ronald@epa.gov

ABSTRACT

Considerable concern has been raised regarding research reproducibility both within and outside the scientific community. Several factors possibly contribute to a lack of reproducibility, including a failure to adequately employ statistical considerations during study design, bias in sample selection or subject recruitment, errors in developing data inclusion/exclusion criteria, and flawed statistical analysis. To address some of these issues, several publishers have developed checklists that authors must complete. Others have either enhanced statistical expertise on existing editorial boards, or formed distinct statistics editorial boards. Although the U.S. Environmental Protection Agency, Office of Research and Development, already has a strong Quality Assurance Program, an initiative was undertaken to further strengthen statistics consideration and other factors in study design and also to ensure these same factors are evaluated during the review and approval of study protocols. To raise awareness of the importance of statistical issues and provide a forum for robust discussion, a Community of Practice for Statistics was formed in January 2014. In addition, three working groups were established to develop a series of questions or criteria that should be considered when designing or reviewing experimental, observational, or modeling focused research. This article describes the process used to develop these study design guidance documents, their contents, how they are being employed by the Agency's research enterprise, and expected benefits to Agency science. The process and guidance documents presented here may be of utility for any research enterprise interested in enhancing the reproducibility of its science.

There is increasing concern about research reproducibility, and this concern is not limited to one field of study, or to just the scientific literature. In 2011, *Science* published a series of articles on the need to improve reproducibility in a variety of disciplines, including 'omics' (Ioannidis and Khoury, 2011), computational science (Peng, 2011) and field biology (Ryan, 2011). A series of similarly focused articles has been published by *Nature*. In studies specifically designed to evaluate reproducibility within particular research areas, Prinz *et al.* (2011) reported on a Bayer HealthCare study in which, of 67 target validation projects in oncology, women's health and cardiovascular disease, only 20–25% yielded data completely consistent with published literature. Begley and Ellis (2012) also reported being able to reproduce the findings in only 6 of 53 'landmark' preclinical oncology studies. Concerns regarding data reproducibility also have been aired in the nonscientific literature, eg, 'Unreliable Research: Trouble at the Lab' (Anonymous, 2013b). Importantly, none of the nonreproducible data reported by Prinz *et al.* (2011) or Begley and Ellis (2012) was due to fraud. Rather, suspected major reasons underlying the lack of reproducibility were the lack of appropriate statistical considerations, including inadequate attention to power, failure to consider bias during study design, less than rigorous or simply flawed data analysis, and failure to include adequate methods descriptions (Begley and Ellis, 2012; Prinz *et al.*, 2011). Federal funding institutions are responding to concerns regarding data reproducibility and transparency by initiatives such as bolstering training in study design and revision of review processes (Collins and Tabak, 2014). Publishers are modifying editorial policies and practices. The latter has involved abolishing the strict limitations on methods descriptions (Anonymous, 2013a), creating separate statistics editorial boards (McNutt, 2014), providing guidelines on the approaches that should be included (Miller, 2014; Schmidt, 2014), encouraging the use of supplemental materials to describe complex and lengthy methods, and providing for online repositories for step-by-step protocols that are linked to specific manuscripts (Nature's Protocol Exchange, http://www.nature.com/protocolexchange/).

The Office of Research and Development (ORD) is the principal organization within the Environmental Protection Agency (EPA) responsible for research underpinning the public's trust in the Agency's ability to safeguard the environment and human health and well-being. All parties and stakeholders affected by an Agency decision, rule, or guidance, and indeed all U.S. taxpayers, have a strong and legitimate interest in ensuring that the underlying scientific research is well-designed, sufficiently powered, technically sound, properly analyzed, and its interpretations authoritative. The quality and reproducibility of ORD research are critical to the Agency's regulatory decisions being accepted as credible and valid by stakeholders, congress, and the federal judiciary. Other agencies, organizations, and countries adopt the content of Agency guidelines; again, such adoption and use depend on the science foundation underlying the guidelines. Because of these dependencies, EPA has a longstanding quality assurance program in place to ensure reproducibility through assessment and evaluation (www.epa.gov/quality/).

Reproducibility and transparency in today's science increasingly demand acknowledgment and action by the scientific community. This article describes several ORD science reproducibility initiatives framed in terms of heightened global statistical and data transparency drivers, but focuses on the development of study design guidance documents that emphasize statistical considerations and modeling best practices. To the authors' knowledge, this is the first example of a set of guidance documents that are designed to prospectively add value during the design phase for three major research approaches, ie, experimental, observational, and modeling. The complementary nature of the documents also supports their joint use in planning studies that combine these research approaches. The documents also support managerial and quality assurance reviews prior to research implementation, arguably an efficient process for effectively directing resources and research potential. Further, the authors believe this is the first article to interpret recent changes in science from the perspective of their exacerbating effects on data reproducibility (see below, *Global Drivers for Reproducibility and Transparency*). The article concludes by highlighting translation of these initiatives in ORD research, although this article and the study design guidance documents should be of interest to anyone involved in toxicological research, whether in the industrial, government, or academic sectors.

Global Drivers for Reproducibility and Transparency

For today's science, ensuring reproducibility presents more challenges than ever. Such challenges arise in many fields of science and at various stages of a project, from planning and design, through analysis, and finally, publication and dissemination.

Appropriate statistical input during study design and planning is an investment in scientific reproducibility. Design and analysis best practices, long been held to support reproducible research, are gaining renewed prominence. Often, today's studies are focused on problems wherein the inherent effect size is small. The ability to detect true rather than spurious differences in studies of smaller effect sizes is undermined by insufficient power resulting in a corresponding elevated risk of both Type I and Type II errors (Gelman and Weakliem, 2009; Salman *et al.*, 2014). Addressing this problem has seen increased emphasis on the utility of power analysis with its estimation of sufficient sample size for a predefined effect size and variability (Couzin-Frankel, 2013; Schmidt, 2014). Randomization and blinding, two further design elements, hold the potential to reduce bias and are experiencing resurgences in popularity; however, they are not yet routine in animal research (Couzin-Frankel, 2013; Tilson and Schroeder, 2013). Other elements of study design and planning, such as compliance with reporting guidelines and definition of inclusion and exclusion criteria prior to data acquisition, can reduce investigator bias and increase the effect-to-bias ratio, thereby increasing confidence in the generalizability of the results (Fisher *et al.*, 2009; Ioannidis *et al.*, 2014; Macleod *et al.*, 2009).

A relatively new study planning idea, although standard for clinical trials, is public preregistration of studies. This idea is gaining broader support as it is seen as promoting research transparency and credibility. Studies may be preregistered, eg, in the Open Science Framework (https://osf.io) database, to pair exploratory studies with replication studies for subsequent joint publication; such paired studies replace the same number of unrelated studies, potentially reducing Type I error (Nuzzo, 2014). Observational studies and clinical trials, in particular, may also be registered prospectively in ClinicalTrials.gov (Dal-Re *et al.*, 2014; Williams *et al.*, 2010).

Study planning best practices also includes considerations of anticipated data analysis to address research questions. Preparing for analysis decisions during study planning also is an investment in reproducible research. The explosion in data generation capability facilitates use of novel approaches to

attack multifactorial problems but carries with it an elevated risk of Type I errors due to artifacts (Anonymous, 2012). The need for multiple comparison adjustment, a priori plans for replication, as well as careful consideration of statistical power grows as researchers investigate more endpoints jointly as a resource- and cost-saving measure and as high-throughput technologies become more prevalent (Benjamini and Hochberg, 1995; Ioannidis and Khoury, 2011; Sham and Purcell, 2014). Increasingly, researchers recommend calibrating P-values or abandoning their use altogether in favor of reporting effect sizes with confidence intervals or use of multiple or Bayesian approaches (Johnson, 2013; Motulsky, 2014; Nuzzo, 2014; Sellke et al., 2001).

Best practices for reproducible modeling-based science includes planning and definition of anticipated use of models, particularly as modeling research is integrated across media (ie, air, water, land, as well as different biological matrices) and across disciplines (Beketov and Liess, 2012; Laniak et al., 2013). More comprehensive quality standards are required as model complexity expands. Careful documentation of software used, model structure, and input parameters are necessary to ensure that model outputs are reproducible. Uncertainty quantification is vital both for interpreting and communicating model results and for optimal decision making (Silver, 2012).

There are growing expectations that science should be conducted more transparently than ever before, and as a result, there is mounting pressure to make research protocols and data public. These challenges to existing research paradigms also place increased emphasis on more thoroughly developed research designs that include data management and data sharing plans. Processes are being developed in response in both the public and private sectors to make protocols and data public. In a 2013 policy memorandum, the Office of Science and Technology Policy in the Executive Office of the President called for Federal agencies with over $100 million in research and development expenditures to develop plans to make research data publicly available to other scientists, regulators, and the general public (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) (see http://www.epa.gov/open/ for the EPA plan). The general scientific community is also calling for more widespread public access to data (Chan et al., 2014; Dal-Re, et al., 2014). Scientific journals including Nature and PLOS are requiring that data, including supporting data, be made available as a condition for publication in their journals so that others are able to replicate the work and that any restrictions on data access be disclosed at the time of manuscript submission. They also are relaxing or eliminating word limits for methods sections and encouraging authors to submit detailed protocols. These demands for data access and transparency require an increased emphasis on the prospective development of infrastructure and processes to facilitate and support data sharing, including consideration of the start-up and maintenance costs of such repositories.

ORD Reproducibility Initiatives

ORD has several reproducibility and transparency initiatives underway, including case studies piloting new processes to publicly release data and publications and adopting systematic data review (Thayer et al., 2014). Two new initiatives highlighted in this article more specifically address statistical considerations during study design and best practices to consider during the design of modeling studies.

ORD Community of Practice for Statistics

In the fall of 2013, ORD leadership encouraged its scientists to create a more systematic process to address statistical considerations in research design and to raise awareness of the statistics-related challenges and opportunities in ORD research. A later Nature editorial reinforced this scientific need in general; 'Too often, statistics is seen as a service to call on where necessary—and usually too late—when, in fact, statisticians should be involved in the early stages of experiment design (Anonymous, 2014)'. In partial response to this need, a statistics community of practice was established to provide a forum for interested individuals to discuss ongoing and planned projects, to apply group thinking to particularly challenging designs, and to discuss innovative methodologies and approaches in both design and data analysis. The ORD Community of Practice for Statistics (SCoP) was formed in early 2014, currently has over 100 members, and meets monthly. The community is further evolving to serve an educational and informal consulting role within ORD.

SCoP Study Design Guidance Documents

In an effort to strengthen the development and required review of study protocols, ORD leadership charged the newly formed SCoP to develop value-added guidance consisting of questions and criteria focused on study design and best practices to enhance science reproducibility. Experimental, observational, and modeling study design work groups were established in response. The three work groups represented a cross-section of ORD organizations and expertise. Draft documents were reviewed by the SCoP as a whole and finalized documents delivered in August 2014. The study design guidance documents have been made available to all investigators, branch chiefs, and quality assurance managers in ORD.

Development of Study Design Guidance Documents

To fully understand and utilize the SCoP study design guidance documents, it is helpful to know the drivers behind each document and the rationale for why certain directions were taken during their development. Experimental and observational research approaches are foundational for much of ORD's science. The experimental and observational guidance documents began from a common set of well-defined questions developed from discussions in the SCoP meetings. The experimental group adapted them to a focus on classical design of experiments where treatments and other study characteristics are under the control of the investigator. The observational group adapted the initial set of questions to focus on ecological, exposure assessment, and epidemiological studies where the investigator does not have the same level of control as in laboratory-based experiments. Given the wide range of statistical knowledge and skills among the projected users, the experimental work group adopted a decision tree format for the questions to help guide investigators and reviewers in understanding whether or not a topic area would be applicable for a particular study. The observational work group adopted an open-ended question format to promote critical thinking by the investigators and reviewers. Parallel efforts by the two groups led to inserting links to internet sites with illustrations and definitions, refining terminology to standard vocabularies, and balancing the level of detail throughout. The observational and experimental study design guidance documents were finalized by standardizing the content common to both documents and reformatting the experimental guidance document as open-ended questions. The main

distinctions in content are that the experimental document includes randomization and blinding, and the observational document includes population selection and sampling details. The experimental guidance document, Supplementary Table S1, has similarities to Nature Publishing Group's Reporting Checklist for Life Sciences Articles (http://www.nature.com/authors/policies/checklist.pdf) and other previously published checklists (Hooijmans *et al.*, 2010; Hooijmans and Ritskes-Hoitinga, 2013; Landis *et al.*, 2012). The observational guidance document, Supplementary Table S2, has similarities to previously published evaluative instruments and checklists (LaKind *et al.*, 2014; Stroup *et al.*, 2000; Vandenbroucke *et al.*, 2007). A principal distinction is the point of application; the ORD study design guidance documents are intended for prospective use, while previously published checklists were designed for retrospective use.

Modeling research has a long history within the EPA and is an integral component of ORD science, simulating and predicting human health and ecological effects, increasing understanding and knowledge, identifying knowledge gaps, and informing decisions and policies. The reach of modeling is expanding. For example, *in silico* biology increasingly is being used as a tool to augment *in vivo* and *in vitro* research. The modeling work group acknowledged a great deal of material exists to frame the structure of a modeling study and to pose questions for evaluating a modeling proposal. Work group members shared readily available documents, established a reference list, and identified the basic topics expected to be covered in designing the initial development, reuse, or repurposing of a model. The modeling study design guidance document, Supplementary Table S3, has similarities to previously published processes of considerations (McLanahan *et al.*, 2012), yet is set-up as a step-wise set of questions with accompanying guidance to provide context and definitions of terminology.

Resulting Study Design Guidance Documents

The three study design guidance documents, while distinct, overlap for several steps and follow separate threads for others. This section provides general descriptions of these steps, highlighting key similarities and differences (Supplementary Tables S1–S3), perhaps the first time these overall approaches have been compared and contrasted in such a unified manner and for this intent (summarized in Fig. 1 with text in single quotes below referencing text boxes in the flow diagram).

Each study design guidance document first addresses 'Problem formulation, objectives, application', and 'scope' to establish purpose and context for the proposed research. Experimental and observational guidance addresses the key science problems, study objectives, and research questions. The modeling guidance addresses the need for, and intended application of a model, as well as its scope and known limitations. After problem formulation, the observational guidance document addresses the 'Population of interest' noting that careful consideration of the unit of analysis and extrapolation to the population of interest underlies the soundness of research conclusions. The modeling guidance document, in a separate thread, next addresses whether a 'New versus modified model' will be used for the proposed research, recognizing that careful consideration of this topic helps determine the effort needed so that the resulting model is fit for purpose (ie, appropriate for intended use).

Experimental, observational, and modeling study design guidance documents converge for the next step, the research 'Approach' or 'Model design'. Specifically, the experimental and observational guidance documents address the type of study (eg, pilot study vs confirmatory study) and study design (eg, cross-sectional vs time-course), whereas the modeling guidance document addresses the purpose for model development/modification, which then informs model structure, inputs (parameterization), and level of detail, as well as the requirements for evaluation.

Many of the remaining steps in the modeling guidance differ from those in the experimental and observational guidance documents. Specifically, 'Mathematical representation, Input data & parameters, Parameter estimation, *In silico* implementation', and 'Evaluation of predictions, sensitivity & uncertainty analysis' are all unique to the modeling guidance document. 'Mathematical representation' refers to the equations that represent the qualitative structure of a model. 'Input data & parameters' refer to the numerical values (parameters) that describe chemical, biological, or environmental variables; these inputs are incorporated into the model to allow simulation/prediction of the environmental or biological outputs of interest. 'Parameter estimation', or model calibration, refers to the process of adjusting model parameters (eg, input data or estimates used in the absence of data) within a defensible range; the purpose is to refine the model to achieve a desired degree of correspondence between output predictions and actual observations. '*In silico* implementation' refers to the identification and use of hardware, software, algorithms, data inputs, and computational methods, as well as the documentation of model results. Finally, 'Evaluation of predictions, sensitivity & uncertainty analysis' refers to the adequacy (ie, suitability and applicability) of a model to describe the system of interest. Sensitivity and uncertainty analyses are important evaluative procedures, and help establish the credibility of model predictions and identify what improvements need to be made. Predictions may be evaluated by a variety of methods (eg, graphical, cross-validation, comparison of statistical metrics to available 'benchmarks', analysis of residuals, etc.), depending on the context of the model application.

'Sampling details', a research step unique to the observational study design guidance document, addresses methodological information that relates to population sampling (eg, random selection vs cluster sampling) and variable measurement (for primary variables and covariates). All other steps addressed in the experimental and observational guidance documents are common to both areas of research. Specifically, 'Statistical analysis plan' addresses methods for statistical parameter estimation and hypothesis testing, as well as related assumptions, uncertainties, model selection techniques, and methods for presenting/interpreting study results. 'Power analysis' aims to ensure that proposed studies have a sufficient sample size for a predefined effect and assumed power level or sufficient power to detect a predefined effect for a fixed sample size. 'Bias consideration' addresses potential sources of bias to ensure sound planning and robust and transparent procedures for data collection, analysis, and publication. 'Data limitations, multiple comparison adjustments' addresses criteria that may be used to include or exclude samples or individual measurements, methods for addressing confounders and effect modifiers, representation of data below detection or quantification limits, strategies for handling missing or negative measurement data, and issues related to multiple comparisons and techniques proposed for correction. All three guidance documents then converge on the 'Data management & sharing plan'. Data plans are designed to ensure that study data and protocols are adequately
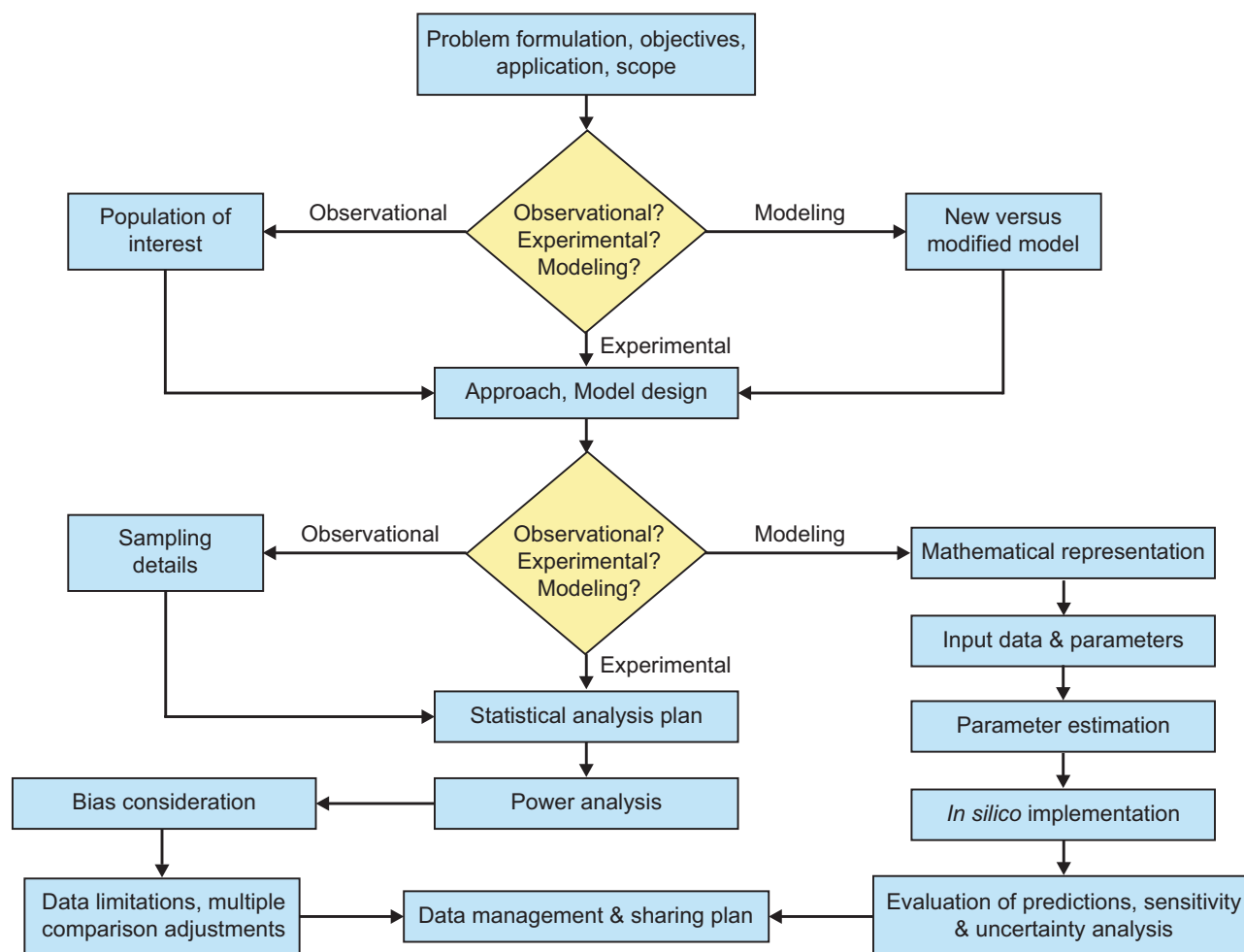
FIG. 1. A Unified Study Design Flow Diagram. The flow diagram depicts common and distinct research design steps based on the experimental, observational, and modeling study design guidance documents.

captured, labeled, stored, and managed during and after completion of this study.

Translation in ORD Research

Implementation steps are underway for use of the study design guidance documents by investigators, branch chiefs, and quality assurance managers in planning and review of research protocols. Meetings are being held with senior and midlevel management, quality assurance managers, and principal investigators to answer questions and gain buy-in for the use of these tools in all ORD programs. Opportunities have been created and others are being sought to critique the suite of guidance documents in terms of their utility, clarity, and usefulness to ORD scientists and managers for study design and review/approval of study plans. These case studies will likely prompt modifications to the guidance documents and, potentially, development of new tools for the program. The case studies will be compiled so that, over time, a library of examples covering a wide range of study designs is available both to ORD scientists, as well as the public. The documents will evolve and mature through use, but an inherent limitation will persist in that checklists and criteria are not surrogates for direct input by statisticians (Anonymous, 2013a).

The SCoP further contributes to ORD's reproducibility initiative through its educational and informal consulting role in

statistical design and best practices. Failure to follow sound statistical design and best practices from research initiation (ie, problem formulation, research planning) cannot be reversed at study end (ie, data interpretation and manuscript preparation). Awareness is the first step and is leading to statisticians (or investigators with extensive statistics training) increasingly being included on research teams at their formation; however, the availability of enough statisticians with the right expertise is an ongoing challenge within and beyond ORD (Anonymous, 2014). There is a recognized need for a bigger statistical presence in science. Interactions between investigators, branch chiefs, quality assurance managers, SCoP statisticians, and scientists will enhance research communication and coordination. These benefits are particularly important as ORD integrates experimental, observational, and modeling research using systems approaches and in a sustainable science framework to address evolving health and environmental effects problems.

As a Federal agency, EPA uses taxpayer dollars for science targeted at preventing public health or environmental harm as well as for science targeting existing public health and environmental damage. ORD seeks continuous improvement in the relevance, quality, and reproducibility of its science. The formation of the SCoP and the development of the study design guidance documents, while not a panacea, represent important steps towards these goals. ORD science is highly regarded and will continue to be so due to the actions being taken in response

to the call to the scientific community to enhance reproducibility. The process used by ORD and the resulting experimental, observational, and modeling guidance documents also may be useful to other organizations in their initiatives to enhance research reproducibility.

SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci. oxfordjournals.org/.

ACKNOWLEDGMENTS

REFERENCES

Anonymous (2012). Error prone. *Nature* **487**, 406.

Anonymous (2013a). Reducing our irreproducibility. *Nature* **496**, 398.

Anonymous (2013b). Trouble at the lab. *Economist* **409**, 26–30.

Anonymous (2014). Number crunch: The correct use of statistics is not just good for science—it is essential. *Nature* **506**, 131–132.

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

Beketov, M. A. and Liess, M. (2012). Ecotoxicology and macroecology—time for integration. *Environ. Pollut.* **162**, 247–254.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B. Met.* **57**, 289–300.

Chan, A. W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gotzsche, P. C., Krumholz, H. M., Ghersi, D., and van der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *Lancet* **383**, 257–266.

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

Couzin-Frankel, J. (2013). When mice mislead. *Science* **342**, 922–925.

Dal-Re, R., Ioannidis, J. P., Bracken, M. B., Buffler, P. A., Chan, A. W., Franco, E. L., La Vecchia, C., and Weiderpass, E. (2014). Making prospective registration of observational research a reality. *Sci. Trans. Med.* **6**, 224 cm1.

Fisher, M., Feuerstein, G., Howells, D. W., Hurn, P. D., Kent, T. A., Savitz, S. I., and Lo, E. H., for the STAIR Group (2009). Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* **40**, 2244–2250.

Gelman, A. and Weakliem, D. (2009). Of beauty, sex and power. *Am. Sci.* **97**, 310–316.

Hooijmans, C. R., Leenaars, M., and Ritskes-Hoitinga, M. (2010). A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. *Altern. Lab. Animals* **38**, 167–182.

Hooijmans, C. R. and Ritskes-Hoitinga, M. (2013). Progress in using systematic reviews of animal studies to improve translational research. *PLoS Med.* **10**, e1001482.

Ioannidis, J. P. and Khoury, M. J. (2011). Improving validation practices in "omics" research. *Science* **334**, 1230–1232.

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., Schulz, K. F., and Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383**, 166–175.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proc. Natl Acad. Sci. U S A.* **110**, 19313–19317.

LaKind, J. S., Sobus, J. R., Goodman, M., Barr, D. B., Furst, P., Albertini, R. J., Arbuckle, T. E., Schoeters, G., Tan, Y. M., Teeguarden, J., Tornero-Velez, R., and Weisel, C. P. (2014). A proposal for assessing study quality: Biomonitoring, Environmental Epidemiology, and Short-lived Chemicals (BEES-C) instrument. *Environ. Int.* **73**, 195–207.

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., Finkelstein, R., Fisher, M., Gendelman, H. E., Golub, R. M., Goudreau, J. L., Gross, R. A., Gubitz, A. K., Hesterlee, S. E., Howells, D. W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S. E., Levine, M. S., Macleod, M. R., McCall, J. M., Moxley, R. T., 3rd, Narasimhan, K., Noble, L. J., Perrin, S., Porter, J. D., Steward, O., Unger, E., Utz, U., and Silberberg, S. D. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187–191.

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., and Hughes, A. (2013). Integrated environmental modeling: a vision and roadmap for the future. *Environ. Modell. Softw.* **39**, 3–23.

Macleod, M. R., Fisher, M., O'Collins, V., Sena, E. S., Dirnagl, U., Bath, P. M., Buchan, A., van der Worp, H. B., Traystman, R., Minematsu, K., Donnan, G. A., and Howells, D. W. (2009). Good laboratory practice: preventing introduction of bias at the bench. *Stroke* **40**, e50–e52.

McLanahan, E. D., El-Masri, H. A., Sweeney, L. M., Kopylev, L. Y., Clewell, H. J., Wambaugh, J. F., and Schlosser, P. M. (2012). Physiologically based pharmacokinetic model use in risk assessment-why being published is not enough. *Toxicol. Sci.* **126**, 5–15.

McNutt, M. (2014). Raising the bar. *Science* **345**, 9.

Miller, G. W. (2014). Improving reproducibility in toxicology. *Toxicol. Sci.* **139**, 1–3.

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *J. Pharmacol. Exp. Therap.* **351**, 200–205.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* **506**, 150–152.

Peng, R. D. (2011). Reproducible research in computational science. *Science* **334**, 1226–1227.

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**, 712–713.

Ryan, M. J. (2011). Replication in field biology: the case of the frog-eating bat. *Science* **334**, 1229–1230.

Salman, R. A., Beller, E., Kagan, J., Hemminki, E., Phillips, R. S., Savulescu, J., Macleod, M., Wisely, J., and Chalmers, I. (2014). Increasing value and reducing waste in biomedical research regulation and management. *Lancet* **383**, 176–185.

Schmidt, C. W. (2014). Research wranglers: initiatives to improve reproducibility of study findings. *Environ. Health Perspect.* **122**, A188–A191.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Am. Stat.* **55**, 62–71.

Sham, P. C. and Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346.

Silver, N. (2012). *The Signal and The Noise: Why Most Predictions Fail—But Some Don't*. Penguin Press, New York.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A., and Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group*. JAMA **283**, 2008–2012.

Thayer, K. A., Wolfe, M. S., Rooney, A. A., Boyles, A. L., Bucher, J. R., and Birnbaum, L. S. (2014). Intersection of Systematic Review Methodology with the NIH Reproducibility Initiative. *Environ. Health Perspect.* **122**, A176–A177.

Tilson, H. A. and Schroeder, J. C. (2013). Reporting of results from animal studies. *Environ. Health Perspect.* **121**, A320–A321.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gotzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J. J., Egger, M., and STROBE Initiative. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology* **18**, 805–835.

Williams, R. J., Tse, T., Harlan, W. R., and Zarin, D. A. (2010). Registration of observational studies: is it time? *Can. Med. Assoc. J.* **182**, 1638–1642.