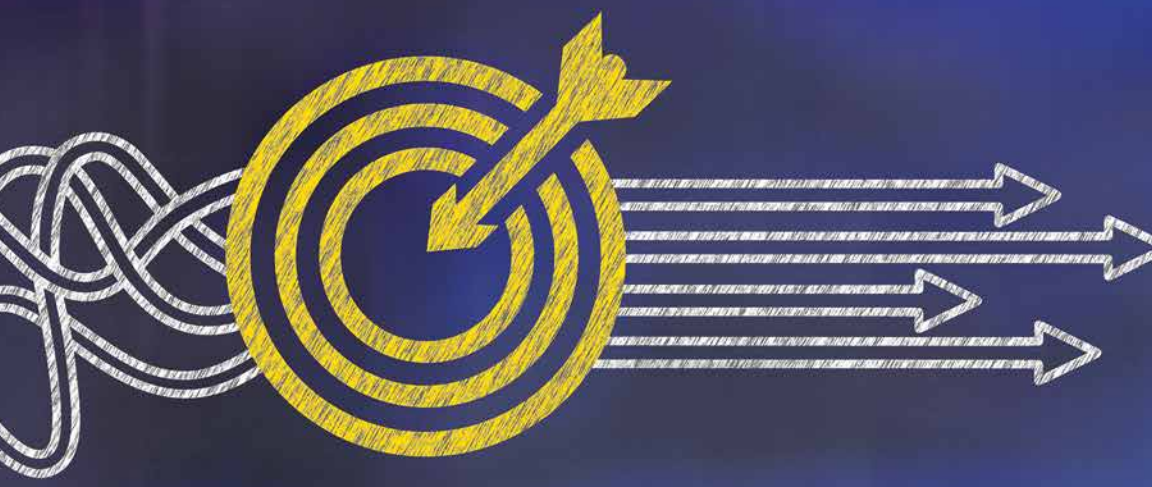
Reproducibility

A Primer on Semantics and Implications for Research



Edo Pellizzari, Kathleen Lohr, Alan Blatecky, and Darryl Creel

Reproducibility: A Primer on Semantics and Implications for Research

Edo Pellizzari, Kathleen Lohr, Alan Blatecky, and Darryl Creel

RTI Press

RTI INTERNATIONAL

Library of Congress Control Number: 2017951510

ISBN 978-1-934831-21-2
(refers to paperback version)

RTI Press publication No. BK-0020-1708

https://doi.org/10.3768/rtipress.2017.bk.0020.1708

www.rti.org/rtipress

Cover design: Alisa Clifford

The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

This publication is part of the RTI Press Book series.

RTI International
3040 East Cornwallis Road, PO Box 12194
Research Triangle Park, NC 27709-2194, USA
www.rti.org/rtipress
www.rti.org

Contents

(continued)

Tables

Figures

Preface

Reproducibility is a core tenet for conducting and validating scientific investigations and thus producing high-quality research. The reproducibility of a result or finding is a fundamental goal in science. Often, investigators, publishers of journals, and others give low priority to results or findings that are merely confirmatory of previous findings. Increasingly, however, this narrow focus on "new" research results is seen as a challenge to scientific progress and to the acceptance of research (or science and the scientific method) by the public at large. This growing acknowledgment of the importance of reproducibility has been accompanied by rising confusion about this and related concepts, such as repeatability, replicability, transparency, and quality. Globally, experts are pointing to the muddled state of affairs in terms of understanding the literal and technical meanings of these constructs—and using them correctly. Conflating the terms undermines communication across scientific disciplines and endeavors and makes the practical processes of ensuring rigorous multidisciplinary research more difficult.

We consider the issues about reproducibility to be core elements of high-quality research. Most research organizations have formal expectations and policies or programs setting standards for high-quality research. For example, RTI's commitment to robust research is articulated in its own quality manual. Its goal is to ensure excellence for all RTI's products and services, from project conception through completion and delivery of research data and final reports, as well as for publications in peer-reviewed journals, the RTI Press, and other venues. The challenges, and the motivation for this publication, lie more in understanding concepts around reproducibility—their similarities and differences, their nuances, and their applications in everyday scientific activities.

Late in 2015, RTI International's Fellow Program lent its support for a small group of Fellows to explore issues relating to reproducibility, replication, transparency, and rigor in the research enterprise. The group regarded "research" as encompassing a full range of scientific endeavor, technical services and assistance, and similar work done across RTI. A

smaller set of this group developed a background report, which eventually became this monograph. Appendices to this monograph highlight some practices and resources that are pertinent to the broader requirements of reproducible or transparent research.

To help sort through this confusing environment and offer a common understanding and lexicon across disciplines, we have developed this primer. We hope that it will advance the nation's portfolio of scientific research—for those engaged in the laboratory, natural, clinical, and social sciences—and ensure that users can address the issues of reproducibility and rigor with confidence and conviction.

—Edo Pellizzari, Kathleen Lohr, Alan Blatecky, and Darryl Creel
August 2017

CHAPTER 1

Introduction

"Non-reproducible single occurrences are of no significance to science."

—Karl Popper, 1959

Background and Context

Science is allegedly in the midst of a reproducibility crisis,[1] but questions of reproducibility and related principles date back nearly 80 years.[2] Further, since at least the beginning of the 2010s, numerous controversies in a variety of disciplines have arisen stemming from failure to reproduce studies or their findings, and reproducible research has become a critical issue in scientific discourse.[3] Scientists across numerous disciplines have expressed serious concerns and engaged in both philosophical and practical discussions about reproducible research. Disciplines represented in these discussions include biology,[4] biomedical[5] and preclinical research,[6,7] business and organizational studies,[8-12] computational sciences,[13,14] drug discovery,[15] economics,[16,17] education,[18-22] epidemiology and statistics,[23-25] genetics,[26] immunology,[27] policy research,[28] political science,[29-36] psychology,[29,37-43] and sociology.[44]

As a case in point: research indicates that more than half of psychology studies fail reproducibility tests.[41,45] A team of highly regarded researchers, many of them authors of the studies under review, attempted to reproduce 100 experimental and correlational studies published in three psychology journals[41,45] using the reported designs and original materials when available. The results: (1) 62 percent of the replication attempts failed, (2) the reported replication effects sizes were approximately half the magnitude of the original effects, and (3) although 97 percent of the original studies reported statistically significant results at the 95 percent confidence threshold, only 36 percent of replicated studies found significant results at that same threshold.[41,45]

Subsequently, a team of psychologists led by Gilbert and colleagues claimed that their analysis, which tried to reproduce what the Open Science Collaboration had done, completely invalidated the pessimistic conclusions that had been drawn from the landmark study.[46,47] The disagreements between

the conclusions of these two major studies stem from differences in statistical methods used and interpretations of statistical results obtained.

The growing chorus of concern, from scientists and laypeople alike, is that reproducibility of biomedical research is failing.[48,49] The US National Institutes of Health (NIH) shares this concern. Collins and Tabak, for example, cite several problems.[5] These are chiefly deficits in experimental design that ignore fundamental properties of statistical theory, which may occur in part because of improper or inadequate training, incomplete documentation of experimental design, and a need to maintain a "competitive edge" by making provocative statements rather than presenting technical details about the research in question.[5] To these concerns one might add the lack of incentives, time, and resources, which hinders many scientists from trying to reproduce earlier work. All these issues have contributed to the lack of reproducibility in the biomedical sciences.[5]

Similarly, computational science (sometimes rendered as scientific computing) faces a credibility crisis.[14,50] Computational science is a rapidly growing, multidisciplinary field that uses sophisticated computer capabilities (hardware, software, and advanced analytic techniques or modeling) to solve complicated problems in the biological, physical, and social sciences. Computing results are often presented rather loosely in journal articles, at conferences, and in books.[14] Researchers cannot always verify most of the computational results presented at conferences and in papers today.[14] Too often computations are taken at face value—by both experts and the public at large.

The state of reproducible research was further exposed in a 2016 *Nature* survey of 1,576 researchers. The survey results revealed that "more than 70 [percent] of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."[1] Among those surveyed, 52 percent believed that the crisis is significant, 32 percent believed it was only slight, and 3 percent did not believe a crisis even exists. Furthermore, "73 [percent] of the respondents believe that at least half of the papers in their field can be trusted, with physicists and chemists generally showing the most confidence."[1] Microbiologist Casadevall adroitly described the status of reproducible research: "At the current time there is no consensus on what reproducibility is or should be."[1]

More people have more access to data than ever before, and more of them are attempting to publish their analyses than ever before. The Office of Science

and Technology Policy has been promoting, or even requiring, that federal science agencies plan to make research data available to the public, industry, and the scientific community.[51] US agencies such as the NIH and the National Science Foundation have begun to require that data from all awards be made widely available.[52] When this trend toward open access to data is coupled with the rapid growth and importance of data-intensive research across all the domains of science, many new issues arise and more common questions become more complicated to address. Private-sector foundations also voice concerns.[53]

In particular, use, reuse, and management of data—and of course issues of reproducibility and replication—become increasingly critical. Provenance, governance, and curation of data, personal identifiers, and metadata must be adequately addressed; if they are not, reproducibility or replicability becomes impossible. This is especially the case if scientific agencies do not establish policies to require more detailed information about how their investigators generated and analyzed their data and do not provide resources for reproducing studies.

Moreover, inadequate analytical skills can produce scientific findings that are neither replicable nor reproducible (and have done so).[54] High-profile cases have been reported in several research fields. Examples include the inability to reproduce Reinhart and Rogoff's work on the ratio of debt to gross domestic product[55] and Regnerus' New Family Structures Study of children with same-sex parents.[56,57] The retraction of the LaCour and Green work on transmission of support for gay equality was actually a case of fabrication of data (scientific misconduct),[58] but it was identified through an effort to reproduce the original study and so we include it here. Other types of studies have reported similar reproducibility outcomes.[6,57] Issues of scientific misconduct, such as deliberate fabrication of data and findings, and rising emphasis on issues of both financial and nonfinancial conflicts of interest, are beyond the scope of this monograph.

The fact that, in many cases, the original authors were willing to have their work scrutinized suggests that scientific misconduct was not a reason for failure in reproducibility. What it does indicate is that widespread publication of work, when seen by other investigators after publication, may well not stand up to scrutiny.

A major reason for this outcome has been characterized as unconscious bias.[59] Often, researchers fool themselves, especially in the data analysis phase,

and follow their "instincts." However, when large number of variables are involved in such research, instincts can no longer be trusted.[60] The significant growth of science and continued emphasis on publication both contribute to this dilemma.

Private-sector, academic, and nonprofit groups are leading multiple efforts to reproduce selected published findings. So far, the results do not make happy reading.[61] As indicated at the beginning of this chapter, several high-profile endeavors were unable to reproduce the large majority of peer-reviewed studies that they examined. These efforts have taken place in many fields, including biology, biomedical, chemistry, economics, medicine, psychology, and sociology.[1,61] Meanwhile, additional discussions and a recent study by the US National Academies of Sciences, Engineering, and Medicine entitled "Fostering Integrity in Research" highlight the importance of addressing scientific integrity and reproducibility for the nation's research enterprise.[62]

Practical Considerations

Because of ongoing trends among leading peer-reviewed journals, research funders, and others who are addressing these issues, authors increasingly need to pay close attention to reproducibility issues in all their research publications. Such attention begins, ideally, at the outset of the research, but it carries through to final publication of research results.

For example, some journals waive length restrictions or provide for web-based supplemental materials (i.e., appendices) to facilitate reporting methodological details.[5] Others promulgate publication review criteria that specifically consider the ability to replicate or reproduce reported results.[63] Finally, as of late 2016, more than 30 journals, associations, and professional societies, particularly in the biomedical realm, have agreed to endorse principles and guidelines for reporting preclinical research.[64] These principles and guidelines cover rigorous statistical analysis, transparency in reporting, sharing of data and materials, and inclusion of image-based data and descriptions of biological material in decisions about manuscript acceptance and publication.[64]

A second consideration is the potentially substantial cost implications of thoroughly planning for and addressing reproducibility issues.[65] Such costs may be based, at least in part, on the degree of scrutiny that is proposed to address reproducibility issues throughout the study and publication of results.

Although pursuing a comprehensive plan may entail considerable work and costs at the outset, the benefits to doing so may be substantial as well.[14]

A third, and possibly knottier problem, is that even though scientists agree about a "crisis" for carrying out reproducible research, they are unable to agree on what "reproducibility" means.[66] The proliferation of related concepts, such as repeatability and replicability (and multiple subsets of these constructs), as well as notions of transparency and documentation or of scientific rigor, robustness, and quality, simply add to the confusion. Misuse and misunderstanding ensue, and communication about the problems becomes confusing, hampering decisions about addressing the issues.

Adding to the problem is a belief, among both experts and the public at large, that scientists cannot or will not adopt a universal (i.e., shared) definition of reproducibility (or the related terms). That is, instead of advocating for commonly agreed-upon definitions of this term, some leading scientists have proposed expanded sets of terms and terms with various distinctions. Whether or how this trend might illuminate the issues and reduce the conceptual and lexicographical muddle remains to be seen.

The complex (not to say chaotic) nature of animal, human, social, and physical systems imposes limitations on reproducing (or repeating, or replicating) scientific experiments and other types of studies. Thus, we acknowledge that high-quality research can be, but may not be, reproducible or replicable.[67] It should, however, always be transparent and fully documented. Finally, it should be robust enough to assure scientists and laypeople alike about the confidence they can have in reported findings and ensuing policies that society may want to adopt.

Purpose of This Monograph

To address these conceptual, definitional, and practical issues, a group established by the RTI International Fellow Program began in 2015 to develop a primer on the topic. Its remit included examining the semantics of the terminology and exploring some key implications for the research enterprise broadly conceived. The main audience is scientists and researchers who are not, at present, steeped in the debate.

The motivation was, in part, a belief that appreciating the nuances of reproducibility may help researchers to communicate about problems of "one-off" study results or of apparently differing findings from studies that purport to repeat earlier investigations. We also aimed to provide some background

and a platform for later documents or educational tools that can examine quality and transparency in the projects being conducted by the research community. We believe that such information in a reference document can help to forestall "non-reproducible" research and promote rigorous, high-quality research. We hope to help scientists and society as a whole have confidence in the integrity of the work and the soundness of published findings. Finally, we sought to present the semantics of reproducibility and reproducible research, because those ideas cross numerous disciplines of research.

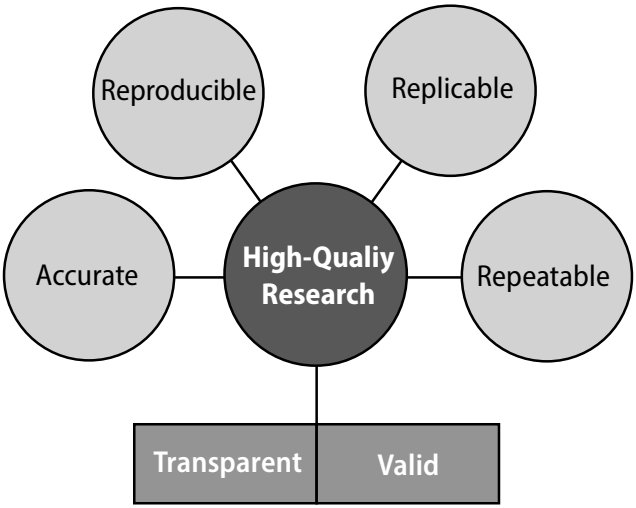
In the next chapter, we define several main concepts and terms relating to reproducible research and show how they differ from (or are consistent or inconsistent with) authoritative definitions across the bench sciences and the social sciences. In the following chapters, we draw out some universal themes that can help scientists from different natural and social sciences backgrounds to communicate clearly, collaborate productively, and generate reliable, valid, and believable information to guide future research and policy. We also address practical steps for assuring transparent, valid, and high-quality research, particularly with respect to publishing such work. Our concluding chapter recaps major points. Appendices A and B cover detailed or technical issues about regulated research and subsets of concepts such as transparency.

Definitions of Concepts and Terms

Introduction

Figure 1 depicts the concepts that we address in this monograph. Our basic premise is that high-quality research is closely associated with concepts such as being accurate, reproducible, replicable, and repeatable—all underpinned by the idea that these scientific endeavors are transparent and valid. Definitions of these concepts are provided in Tables 1 and 2 and the accompanying text.

Figure 1. Key concepts of high-quality research



The most authoritative source that we cite for the definitions of terms is the National Institute of Standards and Technology (NIST; https://www.nist.gov). Across both natural or bench science and the social and policy sciences, NIST offers the most well-known and respected foundation for clarifying and standardizing the lexicon for scientific communication across fields and disciplines. We discuss these key concepts next, relying on NIST when possible; when not, we offer dictionary definitions or cite interpretations or definitions from other scientific experts.

Key Concepts and Related Definitions

For the key concepts presented in Figure 1, Table 1 provides authoritative definitions, some common alternatives that appear to be consistent with NIST definitions, and definitions that are inconsistent with NIST definitions. We rely chiefly on definitions from NIST; others derive from organizations such as the International Union of Pure and Applied Chemistry and other expert groups. We either adopted (i.e., repeat verbatim, identified in Table 1 by quotation marks) or paraphrased definitions from longer discussions in the literature cited at the end of the table. Our goal for the latter step was to make the definitions succinct.

For this monograph, we are using NIST definitions as the standard lexicon. We believe that basing our vocabulary on NIST definitions will be helpful for scientists and researchers working not just in their own fields but also in multidisciplinary teams for whom conflicting definitions can muddle communications.

Of particular importance is the crucial distinction between reproducibility and replicability. The differences lie mainly in whether every element of a given study must be done in exactly the same way in any effort either to do it again or to confirm original results. That requirement is basically the "restrictive" aspect of replicability. By contrast, reproducibility does not turn so critically on such elements as having identical subjects, methods, and other components; that is, at least one component differs from the original research study. Making a clear distinction is important when gauging the credibility of a study's outcome or conclusion because its integrity generally increases from replication to reproduction, when investigators vary one, several, or all components.

Table 1. Definitions and extensions of key concepts for science and research

Key Concepts	Definitions Adapted from the National Institute of Standards and Technology (NIST)	Definitions in Common Parlance and Their Consistency with NIST Definition
Quality	"It is an encompassing term comprising utility, objectivity, and integrity of research."[a] Utility "means disseminated information is useful to its intended users."[a] Objectivity "means the information is accurate, reliable, and unbiased, and that information products are presented in an accurate, clear, complete, and unbiased manner."[a] Integrity "means information is safeguarded from improper access, modification, or destruction."[a]	• An attribute of research that is measured against a formal or informal standard and that reflects the degree of excellence (or lack of it) of that research.[b] *Comment:* Consistent with NIST definition.
Accuracy	A qualitative term referring to the "closeness of the agreement between the result of a measurement and the value of the measurand."[c]	• "The degree to which the result of a measurement, calculation, or specification conforms to the correct value or a standard."[b] • "The quality or state of being correct or precise."[b] Note: Accuracy implies precision or exactness owing to the care with which the information was assembled. *Comment:* Consistent with NIST definition.
Reproducibility	The "closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement" (from an experiment or nonexperimental study using different conditions, such as a different principle of measurement, different method of measurement, different individuals, and different locations).[c]	• The ability to be reproduced or copied.[b] • As reproducibility relates to the quality of an endeavor, it is the "variation in the average measurements of different appraisers who measure the same items using the same measuring equipment."[d] *Comment:* Consistent with NIST definition. "Where **independent** investigators subject the original data to their own analyses and interpretations. Reproducibility calls for data sets and software to be made available for (1) verifying published findings, (2) conducting alternative analyses of the same data, (3) eliminating uninformed criticisms that do not stand up to existing data, and (4) expediting the interchange of ideas among investigators."[e] *Comment:* Consistent with NIST definition.

Bold added for emphasis

continued

Table 1. Definitions and extensions of key concepts for science and research *(continued)*

Key Concepts	Definitions Adapted from the National Institute of Standards and Technology (NIST)	Definitions in Common Parlance and Their Consistency with NIST Definition
Reproducibility *(continued)*		• "The ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator."[f] Consistent with NIST definition if an **independent** investigator repeats the work. Otherwise, if the same investigator repeats the work, then this is the definition for replicability. • "The ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline."[g] *Comment:* Consistent with NIST definition if an **independent** investigator repeats the work. Otherwise, if the same investigator repeats the work, then this is the definition for replicability.
Repeatability	The "closeness of agreement between the results of successive measurements of the same measurand carried out under the same conditions of measurement" (measurement procedure, observer, measuring instrument, and location).[c]	• The closeness of agreement between independent results obtained when the same study is replicated on identical test material or subjects, using the same measurement procedure, same observer, same measuring instrument, the same location and under the same conditions.[h] *Comment:* Consistent with NIST definition.
Replicability	NIST does not formally address replicability. NIST does, however, use the term in specific applications without defining it or distinguishing it from repeatability (see above). Replicability can be (or not be) consistent with repeatability. Another distinction is that replicability is the act of repeating exactly, whereas repeatability is a measure of the closeness of the act. We focus on replicability rather than repeatability because the scientific literature tends to use the former term, but does so in conflicting ways.	• "The ability to obtain an identical result when an experiment is performed under precisely identical conditions."[i] *Comment:* Consistent with implicit NIST definition. • "The ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected."[f] *Comment:* Consistent with implicit NIST definition, assuming it is the original researcher performing the work. • "The chance that an **independent** experiment targeting the same scientific question will produce a consistent result."[g] *Comment:* Inconsistent with implicit NIST definition. This is the definition of reproducibility.

Bold added for emphasis

Table 1. Definitions and extensions of key concepts for science and research *(continued)*

Key Concepts	Definitions Adapted from the National Institute of Standards and Technology (NIST)	Definitions in Common Parlance and Their Consistency with NIST Definition
Replicability *(continued)*		• "Replicated by multiple **independent** investigators using **independent** data, analytical methods, laboratories, and instruments."[e] *Comment:* Inconsistent with implicit NIST definition. This is the definition of reproducibility.
Transparency	"A matter of showing how you obtained the results being disseminated."[a]	• The clarity and completeness of any scientific report, the extent to which intended audiences can readily understand that report, and the extent to which it provides full disclosure and is free of obfuscation or evasion. *Comment:* Consistent with NIST definition.
Validity	NIST does not formally address validity. The concept is not subsumed in any other concept defined by NIST.	• "The quality of being logically or factually sound."[b] • The extent to which a concept, measurement, or conclusion is well-founded and corresponds accurately to the real world.[j] • The ability to achieve accuracy and reproducibility in an experiment or nonexperimental study. • Whether a scientific claim is true and whether data analysis can be trusted. • The degree to which the tool (e.g., a test in education; a battery of patient-reported outcomes; an instrument in chemistry) measures what it claims to measure. • In the area of scientific research design and experimentation, whether a study is able to answer scientifically the questions it is intended to answer.

[a] NIST, 2016[68]
[b] English Oxford Living Dictionaries, 2017[69]
[c] Taylor and Kuyatt, 1994;[70] Joint Committee for Guides in Metrology (JCGM), 2012[71]
[d] MiC Quality, n.d.[72]
[e] Peng et al. 2006[23]
[f] Bollen et al., 2015[73]
[g] Leek and Peng, 2015[74]
[h] McNaught and Wilkinson, 1997[75]
[i] Casadevall and Fang, 2010[27]
[j] Scientific Advisory Council of the Medical Outcomes Trust, 2002[76]

Bold added for emphasis

Table 2 documents more thoroughly terms related to aspects of high-quality research (as in Figure 1) and, in particular, reproducibility. We note when these entries are consistent with NIST definitions for a given concept and when they are not; the latter situation arises because, in many cases, the definition for a specific concept is more relevant for a different concept.

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Reproducibility		
Methodological reproducibility[a]	"Ability to implement, as **exactly as possible**, the experimental and computational procedures, with the same data and tools, to obtain the same results."[a]	*Comment:* Inconsistent with NIST definition. This is the definition of repeatability. *Recommended change:* Ability to implement different experimental procedures, with the same or independent data and tools, to obtain the same results.
Results reproducibility[a]	"The production of corroborating results in a new study [by] having followed the **same** experimental methods."[a]	*Comment:* Inconsistent with NIST definition. This is the definition of repeatability. *Recommended change:* The production of corroborating results in a new study by having followed **different** experimental methods.
Inferential reproducibility[a]	"The making of knowledge claims of similar strength from a study replication or reanalysis."[a]	*Comment:* Inconsistent with NIST definition. This is the definition of repeatability. *Recommended change:* The making of knowledge claims of similar strength between the original and reproduced studies
Empirical reproducibility[b]	When detailed information is provided "for non-computational empirical scientific experiments" and observations.[b]	*Comment:* Consistent with NIST definition, except that this definition mainly addresses enabling an independent investigator to reproduce an experiment or study. *No change*.

Bold added for emphasis

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research *(continued)*

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Reproducibility *(continued)*		
Statistical reproducibility[b]	When detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc.[b]	*Comment:* Consistent with NIST definition, but mainly addresses enabling an independent investigator to reproduce an experiment or study. *No change.*
Computational reproducibility[b]	When detailed information is provided about code, software, hardware, and implementation details.[b]	*Comment:* Consistent with NIST definition, but mainly addresses enabling an independent investigator to reproduce an experiment or study. *No change.*
Analytical reproducibility	When detailed information is provided that enables independent analysts to reproduce the logical, modeling, and computational processes by which the original investigators had drawn their conclusions from a given database.	*Comment:* Consistent with NIST definition. *No change.*
Reproducible research when it satisfies transparency criteria[c]	Providing detailed description of methods, documenting the computer code and software environment, making the analytical data set available to other investigators, and following standard methods of distribution for other investigators to access the software, data, and documentation.[c]	*Comment:* Consistent with NIST definition, but is mainly a subcategory that is contingent on transparency. *No change*.
Repeatability conditions		
Same measurement procedures[d]	Can include patients or populations; interventions; control groups or active comparators; and intermediate or final outcomes. They can also include a variety of study design and analytic statistics.[d]	*Comment:* Consistent with NIST definition, but is mainly an elaboration describing the condition. *No change.*
Same measuring instruments[d]	Can cover a wide range of instruments, from laboratory equipment to patient-reported questionnaires, used under the same conditions.[d]	*Comment:* Consistent with NIST definition, but is mainly an elaboration describing the condition. *No change.*
Same locations[d]	Can include laboratories, field sites, health facilities, educational settings, community settings, and homes.[d]	*Comment:* Consistent with NIST definition, but is mainly an elaboration describing the condition. *No change*.

continued

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research *(continued)*

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Repeatability conditions (continued)		
Repetition over a short period of time or the same measurement periods[d]	Applies particularly to bench science projects, or the same measurement periods for other types of experiments or studies, especially when the results or outcomes have time as a variable.[d]	*Comment:* Consistent with NIST definition, but is mainly an elaboration describing the condition. *No change.*
Replicability		
No additional terms	Not applicable	Not applicable
Categories of quality or related activities		
Quality assurance[e]	In general: Use of planned and systematic activities to ensure that research and services meet client-specific requirements. Audits are conducted to ensure that research staff are following the procedures as prescribed in standard operating procedures and protocols.[e] In health care: "[A] full cycle of activities and systems for maintaining the quality of patient care."[f]	*Comment:* Consistent with NIST definition. Mainly an elaboration describing the condition as it applies to research or to health care. *No change.* In health care, this term is related to quality improvement (see below).
Quality control[e]	Use of operational techniques and activities to monitor work processes and detect and eliminate causes of unsatisfactory performance as the research is being performed.[e]	*Comment:* Consistent with NIST definition. Mainly an elaboration describing the condition. *No change.*
Quality, quality measurement, or quality assessment in health care[f]	In general: Quality is defined in Table 1. In health care: In 1990, the Institute of Medicine defined quality (for health care) as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge."[f] In general: Quality measurement of assessment is the use of tools to measure or quantify processes and outcomes of a given service. In health care: "Quality assessment is the measurement of the technical and interpersonal aspects of health care and the outcomes of that care . . . [and] is expressly a measurement activity."[f]	*Comment:* Consistent with NIST definition. Mainly an elaboration describing the condition as it relates to health care. *No change.*

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research *(continued)*

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Categories of quality or related activities *(continued)*		
Quality improvement; continuous quality improvement[g]	In general: Use of various formal approaches to analyze performance and implement systematic efforts to improve it. In health care: "Quality improvement entails continuous efforts to achieve stable and predictable process results, that is, to reduce process variation and improve the outcomes of these processes both for patients and the health care organization and system."[g]	*Comment:* Consistent with NIST definition. Mainly an elaboration describing the condition as it relates to health care. *No change.*
Categories of transparent research		
Reviewable research[h]	"The descriptions of the research methods can be independently assessed and the results judged credible."[h]	*Comment:* Consistent with NIST definition. *No change.*
Replicable research[h]	"Tools are made available that would allow one to duplicate the results of the research."[h]	*Comment:* Inconsistent with NIST definition. This is related more to a basic definition of reproducible research. *Recommended change:* Reproducible research—Making available tools that would allow independent researchers to reproduce the results of their research.
Confirmable research[h]	"The main conclusions of the research can be attained independently without the use of software provided by the author."[h]	*Comment:* Not applicable comparison to NIST definition. *No change.*
Auditable research[h]	"Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved."[h]	*Comment:* Not applicable comparison to NIST definition. *No change.*

continued

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research *(continued)*

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Categories of transparent research *(continued)*		
Open or reproducible research[h]	"Auditable research made openly available. This comprise[s] well-documented and fully open code and data that are publicly available that would allow [others] to (a) fully audit the [study procedures and computations], (b) replicate and also independently reproduce the results of the research; and (c) extend the results or apply the method to new problems."[h]	*Comment:* Consistent with NIST definition. *No change.*
Verification[h]	"Checking that the computer code [that the original investigators describe] correctly solves the mathematical problem it claims to solve."[h]	*Comment:* Not applicable comparison to NIST definition. *No change.*
Validation[h]	"Checking that the results of a computer simulation agree with experiments or observations of the phenomenon being studied."[h]	*Comment:* Not applicable comparison to NIST definition. *No change.*
Validation[h]	"Checking that the results of a computer simulation agree with experiments or observations of the phenomenon being studied."[h]	*Comment:* Not applicable comparison to NIST definition. *No change.*
Categories of validity		
Content or face validity	Numerous definitions of types of validity are used for measurement.[i,j] "Evidence that the domain of an instrument is appropriate relative to its intended use."	*Comment:* Not applicable comparison to NIST definition (i.e., NIST does not have definitions for these categories of validity). *No change.*
Construct validity (e.g., convergent and discriminant validity)	"Evidence that supports a proposed interpretation of scores based on theoretical implications associated with the constructs being measured."	*Comment:* Not applicable comparison to NIST definition (i.e., NIST does not have definitions for these categories of validity). *No change.*
Criterion (e.g., concurrent or predictive) validity	"Evidence that shows the extent to which scores of the instrument are related to a criterion measure."	*Comment:* Not applicable comparison to NIST definition (i.e., NIST does not have definitions for these categories of validity). *No change.*

Table 2. Taxonomies and new, expanded, or overlapping terms, phrases, or concepts for science and research *(continued)*

Term or Phrase	Available Definitions, Descriptions, or Explanations	Consistency with NIST Definition and Recommended Change if Needed
Categories of validity (continued)		
Internal validity: risk of bias in or quality of a study	Generally, the following types of validity are used to characterize studies: Internal validity: "the extent to which the design and conduct of a study are likely to have prevented bias"[k] or "the extent to which the results of a study are correct for the circumstances being studied"[l] Risk of bias: the risk of "a systematic error or deviation from the truth, in results or inferences."[m] Quality: "the extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error."[n]	*Comment:* Not applicable comparison to NIST definition (i.e., NIST does not have definitions for these categories of validity). *No change.*
External validity: generalizability or applicability of the data from a study to other populations or settings	External validity: "inferences about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes."[o] Applicability: "the extent to which the effects observed in published studies are likely to reflect the expected results when a specific intervention is applied to the population of interest under 'real-world' conditions."[p]	*Comment:* Not applicable comparison to NIST definition (i.e., NIST does not have definitions for these categories of validity). *No change.*

NIST = National Institute of Standards and Technology. For NIST definitions, see Table 1.

[a] Goodman et al., 2016[77]

[b] Stodden, 2014[78]

[c] Peng et al., 2006[23]

[d] Taylor and Kuyatt, 1994;[70] JCGM, 2012;[71] McNaught and Wilkinson, 1997[75]

[e] RTI International, 2013[79]

[f] Lohr, 1990[80]

[g] Knox and Brach, 2013[81]

[h] Stodden et al., 2012.[82] For more details, see Appendix B.

[i] Scientific Advisory Committee of the Medical Outcomes Trust, 2002[76]

[j] Cella et al., 2015[83]

[k] Cochrane Collaboration, 2005[84]

[l] Jüni et al., 2001[85]

[m] Viswanathan et al., 2012[86]

[n] Lohr, 2004[87]

[o] Shadish et al., 2002[88]

[p] Atkins et al., 2010[89]

Explication of Key Concepts

Referring to Figure 1 and the definitions presented in Tables 1 and 2, we use this section to provide a broader description of these same concepts. Entries in Tables 1 and 2 do not exhaust the possible variations (sometimes minor) in definitions that have emerged over the years; the literature for some concepts (such as quality or validity) is immense.

Whereas common jargon is often vague, with intersecting descriptions of terms, technical definitions have traditionally been developed to bring rigor and specificity that permits distinguishing between terms or concepts. Nevertheless, as presented in Tables 1 and 2, contradictions occur between technical terms because development of their definitions has not been extensively coordinated across disciplines.

Here, therefore, for each term, we begin with its common parlance. Following that, we provide an expanded discussion of technical definitions as applied in the physical, natural and social sciences.

Quality

Quality is an attribute of something—for our purposes, research broadly conceived—that is measured against some kind of formal or informal standard and that reflects the degree of excellence (or lack of it) of that something (i.e., research). Generally speaking, *quality* is taken to mean *high quality*. The concept, as applied to health care, similar enterprises (such as education), and manufacturing has a long history. Seminal work on topics such as continuous quality improvement or Six Sigma performance dates from the mid-1990s.[90-101]

Quality can subsume concepts such as the processes or outcomes of an activity. These specifically include formal programs intended to ensure the quality of something, such as research. Such programs may well entail formal quality assurance (QA) or quality control (QC) activities (which are discussed more in Appendix A).

Quality as it is applied to research activities or public programs serving society as a whole may be defined by national governments (the federal government in the United States) or other bodies (both subnational and cross-national). Such definitions are often cited in laws, regulations, or guidelines. Professional societies, for example those for health care practitioners, often define quality as well. Research and educational institutions may also issue "quality" guidelines, expectations, or manuals for their personnel.[79]

For research and service activities that are formally regulated, US agencies and professional or trade associations can and do issue various definitions and standards for quality. Appendix A gives examples of such agencies or organizations and lists key elements of federally regulated research. (Limitations on the scope of our efforts precluded documenting such statutes or regulations outside the United States.)

In some cases, clients supporting nonregulated research projects may also require various kinds of QA systems or activities. Nevertheless, most often investigators who are not conducting research projects for regulatory agencies or purposes are often left to their own discretion on how to implement high-quality research. Not surprisingly, this fact leads to considerable variability in how investigators implement (even similar) studies.

Finally, in the context of reproducible research, quality refers to the internal processes, procedures, and ways of thinking or problem-solving that researchers use in conducting defensible and useful research. The specific expectation is that investigators (and others) can assume that they have produced results that meet two basic criteria: (1) the findings are "correct" or "accurate" (within limits of statistical power, defensible decisions about design, statistics, and interpretations, and measurable limits) and (2) the findings will withstand rigorous scrutiny.

Accuracy

Broadly speaking, accuracy is the condition of being correct or truthful. These terms—accuracy and correctness—are somewhat interchangeable; here, we use accuracy. In common parlance, accuracy implies precision or exactness owing to the care with which the information was assembled. In scientific usage, accuracy refers to the degree to which the result of a measurement, calculation, or specification conforms to a correct value or a well-accepted standard.

For reproducible research, accuracy is an attribute of the investigation (or technical services activity, etc.) that reflects the correctness or truthfulness of both the results reported and the attendant interpretations of those results. The extent to which this is true places an upper bound on the quality of the research.

Reproducibility

In the broadest terms, reproducibility has numerous (related) connotations. Fundamentally, in common parlance it means making a copy, duplicate, or close imitation of something or to cause something to be recreated. In other words, reproducibility is the ability to obtain the same result, observation, or output from an experiment, a device, a system, a process, or an entire study.[102] We return to more specific considerations about reproducible research in this section.

The concept of reproducibility has various technical definitions as well. For instance, in measurement terms it can refer to the extent to which some observation is (or can be) obtained more than once over some (short or long) period of time. (The length of those time periods will differ, perhaps drastically, depending on the effect of time as a variable.) As reproducibility relates to the quality of an endeavor (and especially in considering the quality of health care), it is the "variation in the average measurements of different appraisers who measure the same items using the same measuring equipment."[72] Reproducibility, in the research context, is the ability of a scientific investigation to be repeated or recreated. In principle, the goal is to produce equivalent (similar, identical) results and interpretations.

A simple example may clarify the difference between reproducibility and repeatability (below). Say the length of a board is measured using a yardstick and determined to be 24 inches in length. The board is measured again but instead with a *tape measure* and again found to be 24 inches. The measured length of the board has been determined to be the same but using a *different tool* and thus the measurand (quantity) has been reproduced.

Reproducibility is the closeness of agreement between *independent* results obtained when carried out under changed measurement conditions.[70,71] Changed conditions in an experiment or nonexperimental study can include different settings, different principles of measurement, different methods of measurement, by different individuals, and different locations. The primary aim may be simply to verify the original results. Other goals may be to do the work again so as to expand the number of subjects (i.e., increase sample sizes and thus power); this may be important when the entire body of research results is to be the target of, for example, cumulative meta-analysis.

In some fields, original research may show little or no effects (i.e., few if any differences between groups, interventions, or other program components). Investigators and funders may not want to "reproduce" the work in this sense

at all. However, for transparency and documentation, all audiences should be able to understand how the original work was accomplished and what, from that, seems to merit being retested in the same or similar circumstances.

Scientists seem to conflate the ideas of *reproducible* and *correct*, but they are not the same thing. A study can be reproducible and still be wrong. (For that matter, a study may be correct yet still not be fully reproducible because of design or other factors.) Nevertheless, scientists generally argue that research that can be independently reproduced is more reliable than research that cannot.

Repeatability

Repeatability is the closeness of agreement between *independent* results obtained when the same method is replicated on identical test material or subjects and under the same conditions.[70,71] As with the example above, a real-world illustration may help clarify differences in these concepts. In the above example, repeatability would apply to multiple measurements using a yardstick or multiple measurements using a tape measure. Such repeated measures yield information about the extent to which results agree when they are produced by measurements made using *the same tool*.[70,71] The variation between measurements (whether replicated or reproduced) can be expressed, for example, by a standard deviation.[70,71,75]

Replicability

Complicating understanding reproducibility is the notion of replicability. The *act* of making a copy or fully reproducing any one or a combination of components is referred to as *replicating* it. This term is not precisely interchangeable with reproducibility, but they are related. Among natural and physical scientists, for example, reproducibility refers to achieving the same phenomenon or result when experimental conditions or tools used are varied to some degree. By contrast, replicability is a more rigorous concept, in that it refers to the ability to obtain an identical result when an experiment is performed under precisely identical conditions and using the same tools.

Referring to the simple example above: Both methods determine that the board is 24 inches long, but reproducing the results of measurement using a different tool lends more credence to the claim (or finding) that the board is 24 inches than does simply replicating it using the same tool. Reproducing the finding implies some greater sense of accuracy to the measurement, even though the measurement may not be accurate in either case.

Transparency

In ordinary parlance, *transparent* or *transparency* has numerous informal and technical definitions. For our purposes, the concept has three main elements: the clarity and completeness of any report or information about a given topic; the extent to which the report or publication can be readily understood (at least by intended audiences); and the extent to which it provides full disclosure and is free of obfuscation or evasion.

Transparency is achieved through accurately and adequately communicating what was done. In the research context, this is typically done through a journal publication or technical report (perhaps including online supplemental materials). Thus, transparency can be seen as an *enabler* for achieving reproducible research.

More specifically, in the context of reproducible research, transparency means that all reports of the investigations are complete, accurate, understandable, and fully documented. The particulars for documentation may differ by scientific field. For instance, ensuring transparency may entail thorough descriptions of cell lines, animals, or antibodies and other reagents used in biomedical experiments. As another example, transparency requires documenting the steps taken in randomizing (or masking/blinding) subjects, researchers, and outcomes assessors in animal studies, experimental (e.g., clinical) trials, and the like. Federally regulated research, examined more fully in Appendix A, places great emphasis on transparency. Transparency is discussed further in later sections of this monograph, and Appendix B provides more details.

Researchers should be clear about their work, for example, by presenting (or making available) tables of the data that lie behind their graphs and figures. Moreover, investigators may need to make the source data available with easy access for other investigators or QA auditors to use or critique.

Validity

As with all the terms examined in this paper, validity has several connotations. For our purposes, it can be defined as the extent to which a concept, measurement, or conclusion is well-founded and corresponds accurately to the real world.[76,83] The validity of a measurement tool (e.g., a test in education, a battery of patient-reported outcomes; an instrument in chemistry) is considered to be the degree to which the tool measures what it claims to measure. In the area of scientific research design and experimentation, validity

refers to whether a study is able to answer scientifically the questions it is intended to answer. In ordinary parlance, the question of validity turns on whether a scientific claim is true and whether data analysis can be trusted.

Concluding Thoughts

The independent genesis of quality-related terms at institutions has led to conflicting definitions across disciplines. Adopting definitions from an authoritative source such as NIST and incorporating them into training curricula at academic institutions and promulgating them through, for example, the lexicons that professional societies use are important next steps. Insofar as many academic institutions that train the next generation of researchers do not deliberately educate graduate students and newly minted investigators in the issues of inconsistent use of definitions as discussed earlier, reproducibility problems will take longer to overcome.

CHAPTER 3

Reproducing or Replicating Studies: Key Considerations and Challenges

Scientific evidence is strengthened when important findings are reproduced, replicated, or repeated. As we introduced in Chapter 2, two classes of studies are important to distinguish—namely, reproductions and replications. Investigators trying to do "the same" study need to choose between (a) trying to do the same study (or arrive at the same results) under changed conditions—i.e., reproduction, and (b) trying to use the same components of the original study under conditions that truly mimic the original work—i.e., replication. As numerous observers have noted, keeping the distinctions clear is more difficult when people use cognates or different types of words or different parts of speech (nouns, verbs, nominalizations) for the concept they are trying to discuss.

The focus needs to be on the outcomes or results of the measurements or interventions of the original study or trial in question. As the complexity of the system studied increases, the ability to repeat a given study of that system generally decreases. In such cases, reproducibility plays a greater role. (Repeatability is a research effort allied more closely with replicability than reproducibility, and we do not address it further here.) The rest of this chapter examines core elements of work to try to reproduce, if not replicate, scientific endeavors.

Reproducibility and Replicability

Reproducibility

As noted in the previous chapter, reproducibility is the closeness of the agreement between results for a given variable coming from the original work and those coming from assessments of that variable under changed conditions of measurement.[27,70,71,75] The later study might be done following the same method on identical test material or types of participants in trials but with

changes in other study design elements such as additional doses of medications or different types of environments or locations).[70,71,75]

Reproducibility is a phenomenon that can be predicted to occur when experimental conditions may differ to some degree; that is to say, reproducibility intrinsically requires changes in some elements of the original investigation. If the research cannot be repeated with precisely the same subjects, then it cannot be replicated but it can be reproduced to determine whether the later study yields the same results or not. Thus, a valid statement that describes a second (or later) study attempting to reproduce the first requires that subsequent investigators specify which conditions of the experiment or study they have changed.

In some social sciences or health research (particularly clinical trials and systematic reviews of such evidence), the construct "PICOTS" can be applied to understand the parameters of the study. PICOTS refers to *populations* (patients), *interventions*, *comparators* (which could include placebos or active interventions of various sorts), *outcomes* measured, *time frames* (for follow-up measurements at the end of the intervention or later follow-up), and *settings*.[103,104] This framework is widely used in evidence-based health care practice work (e.g., systematic reviews[105] and their protocols[106,107]) and can be applied broadly to much social science and clinical research. Such specificity facilitates designing or describing the conditions for studies attempting to either replicate or reproduce the original investigation. Questions related to PICOTS are listed in Table 3.

Table 3. Core components related to the reproducibility of scientific investigations: A PICOTS framework

What is the chief aim of the study?
What are the main characteristics of the trial or study design?
Who or what are the populations or subjects of interest?
On whom or what are such measurements made?
What are its measurement or data collection tasks?
What measurement instruments are used and by whom are they administered?
What reference standards or population norms for instruments or questionnaires apply?
What are the conditions under which study measurements are done?
When are measurement(s) made?
What is the location or setting of the experiment or study?

Conditions relating to reproducibility differ based on the type of research. Generally, however, they may involve any or all of the following components of an experiment or observational study:

What is the fundamental principle (mode) of measurement? For example, to measure the structure of a chemical, investigators might start first using a mass spectrometer with electron bombardment as the principle of measurement. In a subsequent investigation, they (or others) may measure the structure of the same chemical based on infrared light using an infrared spectrophotometer. Other examples of data collection tasks include questionnaires on health-related quality of life, alternative versions of a standardized test that educational programs use, and outcomes assessed for a community-based domestic violence initiative—all of which could be modified in various ways.

Who makes the measurement? The individuals assessing the intervention or comparator variables and outcomes may be investigators, independent outcomes assessors, subjects or patients, administrators of settings, or various other persons (equipment operators, technicians, proxies for children, and so on). The qualifications for different individuals conducting the measurement may vary, possibly affecting the reproducibility of the outcome.

What measurement instrument(s) do the investigators use? These can range across many types of hardware, devices, data collection forms, and questionnaires of considerable variety. Moreover, instruments can be administered, as suggested above, by all kinds of persons involved in the study, including the subjects themselves.

What reference standards or population norms apply? These may include measurement standards, such as those obtained from NIST for calibrating a laboratory instrument. They can also be thresholds for (normal vs. abnormal) diagnostic tests or cut points for results from questionnaires or other instruments completed by schoolchildren, patients, or other subjects.

What are the conditions under which studies deploy measurement instruments? For example, investigators may need to take into account factors such as effects of temperature on the performance of an instrument, schedules followed by an educational institution, or the workflow of a private physician practice.

When are measurement(s) made? Timing issues can include baseline measures, a single measurement at the end of an experiment or study, or multiple measurements during a clinical or after it has ended. Factors to

consider include, for example, season of the year, day of the week, and time of day. They can also involve the timing of exposure to an intervention and length of exposure to the intervention.

What is the location or setting of the experiment or study? Settings include, for instance, laboratories; health or social services facilities; schools; community or fraternal organizations; communities, neighborhoods, or villages; homes and other residential facilities; and even prisons and jails.

Finally, reproducibility may be expressed quantitatively in terms of the dispersion characteristics of the results—e.g., as a standard deviation.[70,71,75] Variance and confidence intervals clarify the uncertainty around estimates of outcomes. In comparing original and later findings to determine whether they agree (such that the later work confirms or supports the earlier work), these statistical elements can be very important for accurately interpreting the "reproduced" study and its findings.

Replicability

As defined in an earlier chapter, replication means that a study conducted by the original investigative team develops new independent data, using the same analytical methods, laboratories, and instruments. The main feature of replication is that the primary elements of the later research are precisely similar to the original experiment, trial, or study.[27] The core components of the research (e.g., the PICOTS) of interest are those outlined in Table 3.

Replicability describes the act or ability of obtaining an identical result when an experiment is performed under precisely identical conditions by the original researcher or possibly other scientists. The notion of repeatability, which closely mirrors or can subsume replicability, refers to the closeness of agreement between the results of successive measurements of the same item being measured that are carried out under the same conditions of measurement.[75] As with reproducible research, differences in results of replicated or repeated studies can be expressed quantitatively with dispersion characteristics such as standard deviations.[70,71,75]

Conceptual, Measurement, and Practical Considerations

Natural and physical sciences have a long and rich history of incorporating these concepts into research. They play a central role in the conduct of social sciences research as well. In all domains, however, investigators need to address numerous conceptual and practical aspects of their work.

Social sciences research addresses, but is by no means limited to, health (including food and nutrition), education, economics, criminal and civil justice, housing and urban development, and environmental sciences (including energy). Attempts to reproduce scientific research in these fields are increasing. Researchers trained in several disciplines, such as economics, psychology, political science, business and organizational studies, policy research, and sociology, have been increasingly discussing reproduction (or even replication) in their various research fields (Caren Arbeit, RTI International, personal communication, 2015).

If results from an experiment or study are concordant when conducted by different investigators, then the experiment or study is considered to be successfully reproduced, even if it was never replicated.[108] In contrast, if the results from different investigators differ, myriad possible explanations must be considered. Some of these may relate to differences in methods or protocols. Because of all the possible explanations, which lead to finer distinctions, scientists across the board are more interested in the reproducibility of results than in the precise replication of experiments or studies.[27,108]

The desirability of showing that later investigations produce the same or similar findings (or even discrepant results) leads to the practical question of how many times researchers should try to reproduce a given study before its results (and conclusions stemming from those results) are accepted.[27] Ideally, an experiment or study should be repeated multiple times before it is considered acceptable;[27] in practical terms, this scenario is not likely. (Indeed, no hard-and-fast rules exist for the number of times that an experiment or study should be done over before being accepted.[27]) Nonetheless, the principle remains: reproducing an experiment or study provides assurance that the effect or result cannot be attributed to chance alone and that it is not an artifact of the experiment or study that yielded a one-time event.[27]

Casadevall and Fang explored these issues in greater depth.[27] Even though the ability of an investigator to confirm an experimental or study result is essential to good science, practical and philosophical limits arise in these endeavors. As noted in Chapter 1, many examples across numerous fields show unconfirmed results in clinical studies, biological experiments, microarray results, and research in many other fields. Another lapse is the failure of clinical trials based on promising preclinical studies. Such observations have led experts to question the validity of the requirement for replication or reproduction in science.

Why should this question arise? One answer is that confirming results by reproduction (or replication, if feasible) is likely to be inversely proportional to the number of variables in an experiment or study.[27] Every variable contains a certain degree of error. Because error can propagate both linearly and nonlinearly, one may conclude that the more variables involved, the more errors can be expected. This relationship thus reduces the likelihood of replicability of any complex experiment or study.

Costs associated with highly complex experiments and studies raise another practical barrier to successful reproduction or replication. Some scientific research, especially in areas of public health, involves longitudinal studies that are so large and of such great duration that they could not realistically be reproduced. A case in point is the renowned Framingham Heart Study, which has been ongoing since 1948. In some cases, researchers may try to reproduce or replicate findings from important studies using statistical modeling. This strategy may obviate the barriers posed by high costs of major trials or large observational studies but has its own methodologic limitations.

Replication can perhaps be done with tightly controlled laboratory experiments. It is often impossible, however, when studying the behavior of dynamic, complex systems, for example at the intersection of human health, the natural environment, and technological risks.[61]

When an experiment or study is relatively easily and inexpensively carried out, then it behooves investigators to ascertain the reproducibility or replicability of a result or finding as fully as possible. Some experts argue that the importance of reproducibility increases in proportion to the importance of a result.[27] Arguably, experiments or studies that challenge existing beliefs and assumptions ought to be subjected to greater scrutiny than those fitting within established paradigms.

Barriers to Effective Reproducible Research

Understanding the definitions of the key terms in this area is crucial for scientific discourse and appropriate conduct of research. Equally challenging, and important, is recognizing several critical issues that confront responsible researchers. In the rest of this chapter we highlight several issues in general terms: the multiplicity of terms and the complexity of taxonomies, the great diversity in research endeavors across the natural and social sciences, differing expectations for regulated and nonregulated research, and the resulting patchwork of authoritative guidance about how researchers should proceed.

Multiplicity of Terms and Concepts

Even when used with respect to the scientific method, the term reproducibility apparently has different connotations, depending on its area of application, as we noted in Tables 1 and 2. Users may apply it subjectively or quantitatively; often, multiple usages are found in research. For example, researchers (and end users of that research) may use reproducibility in conjunction with one or all of the following:

- the uncertainty of measurements
- the functionality of an instrument or technique (e.g., an instrument for making many different kinds of measurements; a tool, form, or questionnaire for collecting information)
- the observations associated with a given procedure or intervention (i.e., data
- the results obtained from a given algorithm, analysis, statistical test, or modeling activity.

As emphasized in Chapter 2, we are advocating adoption of the formal definition[70,71] from the National Institute of Standards and Technology (NIST), which can be found in Table 1. We believe this will reduce the confusion and permit researchers across many different disciplines to communicate and collaborate with a shared understanding of reproducibility (and related constructs that are different in important ways).

Moreover, as discussed below, research is extremely diverse in the natural, physical, and social sciences. Here, we limit the discussion to investigations involving human beings; however, in theory the issues extend to studies of nonhuman genera and species. Thus, the types of studies that might be subject to requirements of reproducibility range widely across all the following categories:

- laboratory experiments (e.g., biological, chemical, or physical studies with experimental designs)
- clinical trials (e.g., randomized controlled trials of health care interventions)
- social science trials (e.g., randomized controlled trials of health care, educational interventions, or other programs to address societal issues)

- nonexperimental (observational) studies of many types, including surveys, longitudinal studies of various sorts (e.g., cohort studies with comparison groups), modeling studies, genome studies, studies based on patient or disease registry data, and the like.

Within these large groupings are a vast array of study designs that need to be taken into account. In short, reproducibility can apply to an extremely diverse universe of research and to situations that run from quite simple to quite complex. Moreover, reproducibility may be ascribed to any one of many integral components that constitute an experiment, trial, or other study.

Reliability and Validity: Related Constructs

Reliability. This concept is typically taken to mean the extent or attribute to which an experiment, measuring procedure, test, system or study yields the same results on repeated trials.[76,83] Numerous aspects of reliability can be invoked, depending on the procedure, instrument, or test. For example, test/retest reliability is the degree to which responses or answers on a "test" are consistent over time when investigators expect them to remain stable; this is typically assessed with correlation coefficients between results of those tests administered at two points that are relatively close in time, such as 1 week.

Validity. Broadly, validity takes several forms (which are catalogued differently in different fields). As laid out in Table 2, these elements include content (or face) validity, construct validity (e.g., convergent and discriminant validity), and criterion (e.g., concurrent or predictive) validity. Added to these constructs are internal and external validity—respectively, the risk of bias in a study and the generalizability or applicability of the data from a study to other populations, settings, and the like. In some research fields, one or more of these types of validity can be assessed or tested—for example, against findings from an earlier use of a test or instrument or with respect to whether "later" results demonstrate the ability of a predictive model to have predicted accurately the outcome of a given assessment or study.

Linked relationships. Reliability and validity are often considered together in considering the soundness of research and research findings. In basic measurement terms, reliability effectively puts a ceiling on validity: studies or measurement techniques that are not reliable cannot be valid.

When taken together, reasonable levels of accuracy and reproducibility (or reliability)—of, say, a device or measurement tool, of algorithms and models,

and even of a laboratory, clinical, or social experiment or a nonexperimental study—imply a degree of validity. The relationships can be complex and not straightforward, however. For example, a method or collection of methods used in a study can be accurate but not reproducible, reproducible but not accurate, neither, or both. A method is valid if it is both accurate and reproducible. Likewise, a study is valid if it is both accurate and reproducible.

Additional Reproducibility Concepts

Over the years, scientists have set forth different terms and usages of these concepts. Some are simply alternative terms or phrases, and some are more complex taxonomies for classifying specific types or classes of reproducibility. Table 2 (Chapter 2) presented taxonomies from various experts in the field.

In addition, Goodman et al. recently augmented the word "reproducibility" with three additional descriptors.[77] *Methods reproducibility* is considered to be capturing the experimental and computational procedures, with the same data and tools, to obtain the same results. *Results reproducibility* is defined as producing corroborating results in a study by having followed the same experimental methods; this is essentially the equivalent of replication. Operationally, however, this goal can be elusive because of problems when studies have substantial random error in any result, which renders difficult concluding that results are "the same." Finally, *inferential reproducibility* reflects the idea of drawing qualitatively similar conclusions from, for example, independent replication of a study or reanalysis of the original study. (It is not the same as reproducing results per se.) Such outcomes may not be possible if different investigators choose different analytic techniques that, even with the same underlying data, produce results potentially consistent with different inferences; in addition, researchers may well draw different conclusions from even the "same" results.

Transparency

Other distinctions about reproducibility are relevant to goals of transparency, documentation, and ultimately the rigor and quality of the research in question. Specifically, these involve empiricism, statistics, computation, and analytics.[82,109] We briefly explored some of these concepts in Table 2 (Chapter 2); we elaborate on them here.

Empirical reproducibility refers to the situation in which investigators provide detailed information about the design and conduct of a given study.[82] This classification includes all types of experiments and trials and all types of

observational (nonexperimental) studies. In practice, transparency through making freely available both the details of how the research team collected data and the data themselves is what enables reproducibility in research.[82] Such documentation thus includes details of procedures, instruments, and methods (e.g., descriptions of study design, survey design, laboratory or field experiments, test subjects and materials, and all related facts about the study).

Statistical reproducibility refers to the circumstances in which researchers or authors provide detailed information, independent of expectations about the idea of "empirical reproducibility," that explains and documents their choices of statistical tests, model parameters, threshold values or cut points, and similar components of quantitative analysis.[82]

Computational reproducibility encompasses detailed information that may go beyond the specific research components noted above. This may include descriptions of hardware, software, programming code, and other details of the implementation or analysis activities.[82] Documentation of standard quantitative software (for statistical testing, meta-analysis, modeling, and the like), such as the specific version of a statistical package (and the developer and city and state or country where it is located) are typical pieces of information expected to be recorded. An increasingly pressing and equally important issue is the documentation of researcher-created code.

Analytical reproducibility enables analysts to reproduce the logical, modeling, and computational processes by which the original investigators had drawn their conclusions from a given dataset. It also implies and may require investigators to calculate uncertainties associated with those conclusions. The process may include any of the following steps:

1. Imputing values for missing data.
2. Editing to correct data values deemed to be erroneous (including eliminating entire records).
3. Altering data to protect confidentiality of data subjects, which may result in some statistical limitations.
4. Constructing additional variables (sometimes called composite or derived variables) from other variables, which can include recoding variables.
5. Linking the dataset to additional data.
6. Modeling the data statistically, including variable transformations, interactions, and choice of "tuning parameters" for models. (Examples

may include prior distributions for Bayesian methods; numbers and types of nodes for neural network models; termination thresholds for partition models; time windows within which an outcome must occur for it to be considered for adverse outcomes from pharmaceuticals; and selection of controls.)

7. Citing the software employed, e.g., packages such as R, SAS, or SUDAAN or customized software; this may include providing convergence criteria for iterative algorithms.
8. Specifying thresholds (e.g., $p < 0.05$) and confidence intervals used to assess statistical significance and levels of confidence.

Investigators should also document "analyses not reported" for at least two reasons. First, researchers should address problems associated with multiple testing of hypotheses and model selection. Second, they need to avoid criticisms of reporting bias; these include publication bias generally but also selective outcome reporting bias, which can include cherry-picking results that favor one element of the research or another. Experts developing systematic reviews by seeking information from multiple sources of data (some published, some otherwise public, and some private) know that such reporting bias and selective reporting are relatively common.

In practice, consensus is lacking about the level of detail needed to describe a study. Moreover, publications may not permit or provide the requisite level of detail. Nevertheless, describing the measurement process, the degree of processing of the raw data, and the completeness of the analytic reporting are all important parts of methodological reproducibility.[77]

Regulated and Nonregulated Research: Yet Another Complexity

Federal regulatory requirements have standardized, and long-standing, definitions for research terms such as reproducibility, repeatability, replication, and accuracy (see Appendix A). These are consistent with NIST definitions as articulated in Table 1. Nonregulatory research, by contrast, has no comparable driving force. Consequently, such research endeavors have more variability in the definitions or meaning of terms and concepts across a large array of fields of study.

Different types of scientific research are inherently easier or harder to reproduce. As FitzJohn and colleagues explain, for example:

> At one extreme is analytical mathematical research, which should in many cases allow for straightforward reproduction [based on published equations]. At the other extreme are field-based studies, which may depend on factors that are not under the control of the scientist [e.g., reproducing a before-and-after study of the effects of a hurricane]. The current frontier of reproducibility is somewhere between these two extremes.[110(p1)]

Guidance for Researchers: A Focus on Quality

Experts in the expanding field of reproducible research may focus more on definitional and conceptual classification schemes than on practical advice about how to proceed. The one exception is federally regulated research, for which considerable guidance is available (as noted above), some of which is related to quality (especially QA and QC procedures).

To provide some ideas about how to proceed, we present a general framework for a research experiment or study that involves at least one step of ensuring that the scientific team or others can take reproducibility into account. This framework focuses largely on principles of "quality" as practical processes. It is intended to depict at what point in the process of designing and conducting scientific investigations these various concepts and terms might be applied.

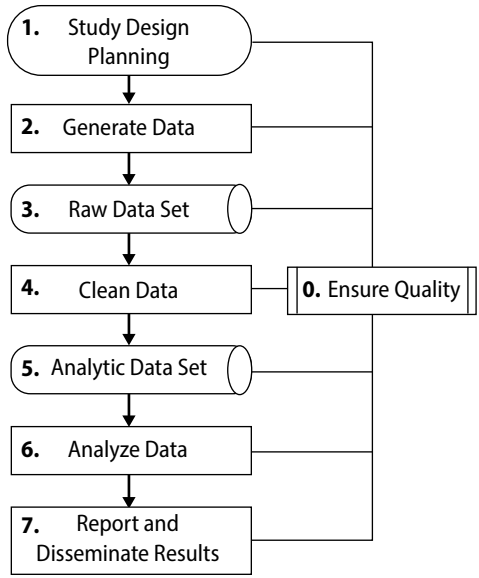
Much of the focus is on what investigators should track, monitor, and report on in publications, final project reports, supplemental materials posted to websites, or other easily accessible venues. The underlying motivations are to foster transparency and strengthen the rigor of the work itself. Thus, this section should be seen as pertinent to both the initial studies and any later (related) ones.

Study Designs and Their Relationship to Reproducibility and Related Principles

Scientists have many different study designs that they can elect to adopt, and each study design has requirements for reproducibility. Some approaches that are common across many designs are critically important for the study to be reproducible.

Figure 2 offers a general framework for study designs—broadly considered to range from experiments and randomized controlled trials to nonexperimental or observational studies. It highlights several conventional design components to illustrate which principles of reproducibility and the

Figure 2. Key steps in experimental or observational research



related constructs (e.g., replicability, but also transparency and rigor) apply. Reproducible research in the context of this framework can refer to a single step, a combination of steps, or the entire process.

The core steps discussed below pertain to both federally regulated and nonregulated research. We discuss some basic properties of these steps below. In some cases, we also note subsidiary activities related to high-quality research (including documentation [see Appendix A] supporting transparency [see Appendix B]). Supporting activities are generally listed as associated with the immediately preceding step so that they have been accomplished before they are needed. In addition, the descriptions of the steps reflect expectations about what investigators have done in documentation and published reports of the research. The remaining element—numbered 0 and discussed first—underscores that efforts to ensure quality affect studies at every step.

Key Steps in Scientific Research

0. Ensuring Quality. Investigators (or other responsible parties) should review each component of the process to ensure that they are producing the highest quality research. This may entail, at a minimum, examining results and the programs in each preceding component to understand the inputs to the next component (see figure). Our view is that researchers should have an overall QA

plan for the study and a specific QC operational plan at every step in the study: planning the study, cleaning raw data, creating analytic data sets, using analytic techniques, generating the data, creating the raw data set, and interpreting data and presenting results. Integrating QC into each task helps to ensure a comprehensive approach to quality. Conducting quality reviews for only some of the components, at the end of the process, or haphazardly will not produce the confidence in the process or product as much as a comprehensive approach to quality can. QA and QC are discussed further in Appendix A.

1. Planning and Designing Studies. Investigators should clearly state the hypotheses of an experiment or the objectives of a nonexperimental (observational) study. They should be specific about any assumptions they make that lead to or determine how the data are to be generated, analyzed, and reported. Two subsidiary elements are particularly important.

Training and validation. If some measures are new to a site or if staff members are inexperienced in administering the instrument, questionnaire, or laboratory technique, then training, reliability testing, validation, and similar tasks may need to be done. These activities help establish the validity of the measurement and the ability of the researcher to collect the measures precisely and accurately.

Curate proven methods. Investigators should explain, describe, register, or cite established protocols, proven code, or high-quality data (or take some or all of these steps), following the general expectations of their respective scientific fields. Curating methods and data supports efforts to make such information and data available to other investigators and thus to make later related (e.g., replication or reproducibility) studies easier to conduct and more accurate.

2. Generating Data. Investigators should clearly report how their data are to be (or were) created. This can involve conducting an assay, making laboratory measurements in a clinical trial, administering a questionnaire (instrument, survey), eliciting observations from third parties (such as family members or teachers), and performing numerous other data collection techniques. Researchers should document comprehensively all methods and procedures associated with data collection; they may provide such descriptions with detailed protocols, manuals of operations, and methods sections of publications (which may include web-based supplemental information in, for instance, peer-reviewed journals).

3. Creating Raw Data Set(s). Data sets come directly from the data generation process. Invariably, some issues arise with such data sets. Examples include missing data (randomly or systematically), contamination of solutions used in experiments, or malfunctions of instruments. Knowledge of these problems will be informative during the data analysis phase, and investigators should document their concerns and the solutions.

4. Cleaning Data. Investigators should develop algorithms to account for the issues they encounter in their raw data sets. Problems can be, for instance, inconsistent, missing, or out-of-range data values. The aim is to prepare the data so that the analytic data sets produced in step 5 have as few challenges or limitations as possible.

Knowledge of the algorithms that investigators developed to deal with problems in the raw data is essential for accurate data analysis. As with the problems themselves, investigators should document the approaches they use to minimize the impact of the problems. Such actions can include making the data logically consistent, imputing missing values, addressing out-of-range or impossible values, and constructing new variables.

5. Creating Analytic Data Set(s). In principle, in steps 3 and 4 investigators will have resolved most, if not all, of the data issues in the raw data sets. The analytic data sets created from the earlier steps process are considered the final data sets for analysis. The investigators' aim is to convey clearly to others what data they intended to analyze and how they obtained the relevant data set(s).

6. Analyzing Data. Investigators should document equations that reflect the conceptual models of the trials or studies, the computer software code used (either uniquely written or off-the-shelf programs), and other tools used to calculate or estimate quantitative or qualitative information. This includes software used in quantitative analyses (e.g., for meta-analysis) and in qualitative analysis (e.g., NVivo, text analysis) that investigators will eventually report in various publications. As with earlier steps, investigators should identify, describe, or otherwise make note of these tools. All analyses conducted need to be documented, even if the results are uninteresting or statistically nonsignificant.

7. Reporting and Disseminating Results. Results comprise both the analytic data and the investigators' interpretations of the information their analyses produced. Of particular importance, investigators must avoid all forms of

publication bias (e.g., actual publication bias by not publishing results at all, outcome reporting bias, or other selection biases).

A Broad View of Attention to Quality in Scientific Investigations

For high-quality research to be implemented effectively, scientists should implement a comprehensive approach to reproducible research during the study design step. A narrow focus on reproducible research may frequently lead investigators (or those responsible for QA and QC) to focus on steps 5 and 6 above—namely, issues about creating analytic data sets and applying analytic techniques and programs. Paying attention to reproducibility and related principles regarding all the components, and specifically steps 3 and 4, may also be valuable. Even if investigators (or QA/QC experts) cannot move all the way back to the data generation step (i.e., step 2), they can still gain valuable knowledge about data by examining the data cleaning step (i.e., step 4). Finally, step 7, where the subject matter, statistical knowledge, and understanding of the scientific, clinical, or policy issues come most into play, is critical in creating truly high-quality research.

Even if investigators manage to execute earlier steps "perfectly," the research can still fail if steps 6 and 7 are not executed as effectively as those earlier stages. In addition to these considerations, transparency (i.e., adequate documentation) in steps 1 and 2 is critical if the research is to be believed and used for clinical or policy decision-making or for independent reproduction or replication.

One of the keys to transparent research is putting the data and computer programs in a state that can be shared fully, as appropriate. This may be peer-reviewed publications, open online sharing, or private sharing if needed based on the provisions of contracts or grants from funders in either the public or private sectors. Furthermore, research teams may well hope that others will use their data, methods, statistical and analytic approaches, and even results in their own research.

Monitoring the use of transparent products is critical not only to show the utility of the work, but also to provide a way to identify issues that the broader scientific, clinical, or policy fields may need to know. When such issues are identified, information about the problems that either the original investigators or later researchers (or both) experienced will allow future investigators to examine methods and protocols to determine where gaps or various mistakes may have occurred. In that way, they can improve their own procedures and move the scientific enterprise ahead.

CHAPTER 4

Transparency in Reproducible Research

Transparency is, at heart, based on the goal of full and accurate documentation of research efforts. The fundamental objective is to enable verification and validation of all steps leading to research results that investigators report in various types of publications or final reports. Achieving this entails meeting certain reporting standards. These may be promulgated by professional societies, journals and journal publishers, and federal agencies when they act as the publisher.

Categories of Research Relating to Transparency

One taxonomy of several categories of research may clarify some issues at the publication stage, drawing on work by Stodden and colleagues.[82] These include (1) reviewable research, (2) replicable research, (3) confirmable research, (4) auditable research, and (5) open or reproducible research. Appendix B explores these points in more detail.

We note below some basic elements for reporting research (for instance, in peer-reviewed journals). We also emphasize the need for making all reports of research conform to readability guidance for correct English, clear writing, and attention to the particular needs of intended audiences.[111] Following that, we give two examples (computational sciences and epidemiology, i.e., a specific research field) of how specific disciplines are turning their attention to these issues.

Fields of Science and Related Reporting Principles

Basic Elements of Reporting Research

The aspects of publishing on research called out below illustrate basic guidance for the types of information typically expected for documentation. In some cases, the specific expectations may be idiosyncratic to a particular field or

type of research; in other cases, the points are applicable across a wide array of investigations. The list below expands on a core set of standards for rigorous reporting studies set forth by stakeholders that the US National Institute of Neurological Disorders and Stroke convened in June 2012.[112] The elements and the guidance cited should not, however, be regarded as a comprehensive list in a field evolving as fast as this one is.

Standards. Authors should follow widely accepted standards for reporting that are pertinent to their particular fields and types of research.[112] These may involve (but are hardly limited to) nomenclature standards; ways to indicate statistical significance (e.g., as confidence intervals rather than simply p-values); rules for using medical terms, abbreviations, and standard or international units; and acceptable acronyms or abbreviations that are idiosyncratic to the natural, physical, or social sciences.

A substantial array of reporting standards has also appeared in the past two decades. One is ARRIVE (Animal Research: Reporting of In Vivo Experiments) for animal studies.[113] For publications reporting on trials of clinical interventions or health care delivery innovations, systematic reviews, or other types of analyses, similar guidance is available—CONSORT,[114] MOOSE,[115] or PRISMA[116,117] (respectively, Consolidated Standards of Reporting Trials; Meta-Analyses and Systematic Reviews of Observational Studies; Preferred Reporting Items for Systematic Reviews and Meta-Analyses). A useful website for numerous reporting standards is the Equator network (www.equator-network.org/) for the main types of research studies seen globally. Finally, some measurement standards assess the reproducibility and accuracy of instruments employed in an experiment or study.

Replications. When a study is a "repeat" by the same research team using absolutely identical methodology, documentation requires that investigators report how often they performed each experiment and whether the results were substantiated by repetition under a range of conditions.[112] Authors must give sufficient information about sample collection to distinguish between independent biological data points and technical replicates.

Sample-size estimation. Authors should be clear as to whether they computed an appropriate sample size when designing their study, including specifying the statistical method of computation.[112] If they did not use a power analysis, then the investigators need to state how they determined their sample size.

Increasingly, the expectation is that investigators reporting on comparative studies will be clear as to whether their power calculations were based on the intent to show superiority vs. inferiority or equivalence.[86,118] In addition, the notion of "optimal information size" is gaining traction.[119]

Inclusion and exclusion criteria. Authors should state clearly the criteria that they used for inclusion and exclusion of any data or subjects and over- or under-coverage of the population with respect to the group from which the subjects were selected. These criteria may be technical metrics (using whatever terms might be appropriate for physical and natural science research). In studies involving human subjects, the criteria may reflect the PICOTS framework mentioned in Chapter 3 and thus might include health conditions, sociodemographic characteristics, and other considerations such as literacy, language, geographic locations, settings, and similar factors.

Randomization. Authors should state whether their samples and study populations were randomized to different arms of an experiment (e.g., one or more intervention groups, one or more groups involving combinations of interventions, and control groups).[112] They must also specify the method of randomization. For nonexperimental studies not requiring (or permitting) randomization, similar descriptions of controls or comparison groups in nonexperimental studies would also be required.

Blinding. Authors need to be clear about whether the investigators were blind to group assignment and outcome assessment.[112] Similarly, they need to state whether patients or participants in studies were masked to their group assignment. If those assessing outcomes are different from the investigators (e.g., clinicians delivering care), whether they also were masked to the assignments of participants to intervention or control groups (or the equivalent in bench experiments) should be documented as well.

Outcomes. As with study subjects or populations, authors should provide a complete list and definition of outcomes of interest for the study. Authors have an obligation to report on the results of all analyses of outcomes specified as part of their methods. They need to avoid all forms of publication and reporting bias. When not all planned outcomes could eventually be measured, then the investigators should explain the reasons in final publications.

Statistics. Statistics must be fully reported in any publication. This includes at least three categories of information: (1) the statistical tests used (including specifying the tests used for different analyses when multiple tests or analytic techniques are applied); (2) the exact value of *N* (for preclinical and clinical studies) or values of *n* (for arms in a trial or groups in observational studies); and (3) the definitions of center, dispersion, and precision measures (e.g., mean, median, standard deviations, standard errors of the mean, and confidence intervals).

Analytic techniques and related software. Investigators have a related obligation to report on all the quantitative, qualitative, or mixed methods approaches that they used for analysis. This is related to statistics (statistical tests), of course, but here the expectations extend to, for instance, direct and indirect meta-analyses, text and content analysis, and the increasing use of approaches to deal with large-scale studies of so-called complex interventions. Related to these items of information is documentation of the software (e.g., name, version, developer city, state or country) that researchers may have used in these analytic steps. This expectation may extend, as well, to libraries, use licenses, and version control.

Results. Authors have an obligation to present all results or to explain what is not being reported or is being reported someplace else. Thus, they need to clarify anything that has been omitted from the reporting for any reason—avoiding publication and outcome reporting biases is critical for audiences to have confidence in the research and the ensuing reporting of that research.[120-123] This is especially important in specific cases: (1) when the results do not support the main hypotheses of the study, (2) when some results are discordant with the main findings, or (3) when some results are statistically significant and others are not at different measurement points.

Computations

With the increasing use of technology to gather, analyze, and report data, reproducible research is becoming more visible in discussions of research methodology. An emerging area of concern, somewhat separate from the basic points in the previous section, is computational science and the credibility of calculations and results reported in various venues. For instance, verifying most of the computational results presented at conferences and in papers today is nearly impossible.[50] Thus, as Donoho states, "computational reproducibility

is not an afterthought—it is something that must be designed into a project from the beginning."[14(p386)]

In this context, analytical reproducibility (defined above) is the ability to recompute data analytic results given an observed dataset and knowledge of the analytic approach and steps.[74] Replicability of a study is the likelihood that an independent experiment targeting the same scientific question (and with no changes to the study parameters) will produce a consistent result. From a computational perspective, according to Leek and Peng,[74] three major components are necessary to achieve reproducibility and replicability. However, we believe that there are actually four components:

1. Making the raw data from the experiment available.
2. Making available the code that transforms the raw data into the analytic data set.
3. Making the statistical code and documentation to reproduce the analysis obtainable.
4. Analyzing the data correctly.

Reproducibility is a minimum expectation that can be directed at any study, insofar as independent investigators have the funds and time to subject the original data to their own analyses and interpretations.[23] Reproducibility calls for data sets and software to be made available for all of the following: verifying published findings, conducting alternative analyses of the same data, eliminating uninformed criticisms that do not stand up to existing data, and expediting the interchange of ideas among investigators. Of paramount importance, however, is protecting the confidentiality and privacy of such data and not breaching any personally identifiable information requirements.

Research-Field-Specific Transparency

Some fields are now tackling reproducibility and transparency in ways specific to their disciplines and the users of their research. For instance, epidemiologists have asserted that a study may become reproducible when it satisfies the transparency criteria (i.e., requirements) noted in Table 4 for four core components of research.[23(p784)]

Table 4. Criteria for reproducible epidemiologic research

Research Component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results reported is made available in a human-readable form. In addition, the software environment necessary to execute that code is available [to other investigator teams].
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar analyses.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

Source: Peng et al., 2006[23]

Given the increasing attention now being paid to accomplishing reproducible research, we expect that additional fields of study, in concert with scientific publishing firms, will develop similar transparency requirements for publishing their research findings.

CHAPTER 5

Conclusions

In this monograph, we have attempted to clarify the broad area of reproducible research, with its attendant (and confusing or even conflicting) lexicon. Our goal is to bring the concerns and issues into greater focus and to make available background information, definitions, and some practical guidance for all readers.

Generally, our conclusions fall into the following categories (implicitly or explicitly). First, researchers must become better educated about these issues in general. Second, they need to grasp the similarities, and particularly the differences, among the concepts and terms, so that they can communicate clearly both within their own fields and, more importantly, across multiple disciplines. Third, scientists also need to embrace these concepts as part of their responsibilities as good stewards of research funding and as providers of credible information for policy decision-making across many areas of public concern. Restoring society's confidence in the scientific enterprise is a crucial part of the responsibility of scientists and researchers across the many challenges facing the global community today. Fourth, a focus on transparency is essential, which means improving our approaches and mechanisms for documenting our work. Fifth, all these considerations are central to ensuring the soundness of the research being pursued and publicized—i.e., attempting to achieve the most rigorous science possible given limitations on time, funding, or other resources.

Finally, we note that this is not a static field, even if experts are converging on standardized definitions and conceptual frameworks. More is needed—as practical guidance—for achieving reproducible research (broadly conceived) and strengthening the transparency and rigor of work across all branches of the natural, laboratory, and social sciences. Responsibility for accomplishing these goals begins with training the next generation of researchers in adopting

a universal lexicon of terms and concepts, applying rigorous study designs, and ensuring transparency in disseminating research findings—as well as encouraging the current scientific community to be more cognizant of issues of reproducibility in this era of rapidly expanding multidisciplinary collaboration and data. Such responsibilities belong to academic institutions, professional societies and associations, and research enterprises, and the obligations extend to funding organizations to support these efforts.

Acknowledgments

The authors are indebted to numerous colleagues associated with the RTI International Fellow Program: Don Bailey, PhD, former chair of the Program; members of the Fellow Program's Scientific Stature Committee, including Eric Otto Johnson, PhD, chair, and Jenny Wiley, PhD, previous chair and Deputy Editor-in-Chief, RTI Press; and participants in the Reproducible Research subcommittee, including Grier Page, PhD, chair, Caren Arbeit, PhD, Philip Cooley, MS, Anthony Hickey, PhD, Meera Viswanathan, PhD, and William Zule, PhD. We are most appreciative of inputs from Alan Karr, PhD, on an earlier manuscript. Joanne Studders, the senior editor for RTI Press, gave us consummate, fastidious service in editing this book, and we are profoundly thankful for her efforts. Finally, we express as well our deep gratitude to Loraine Monroe for exemplary performance in preparing this document.

About the Authors

Alan Blatecky, MBA, MA, ThM, MDiv, is a Visiting Fellow at RTI International; he previously directed the NSF Office of Cyberinfrastructure at the National Science Foundation. Mr. Blatecky is also a member of the Steering Committee for the National Data Service and the US Research Data Alliance. He helped co-found the Renaissance Computing Institute and served as the Deputy Director; he was also the Executive Director of Research and Programs at the San Diego Supercomputing Center, and Vice President of Information Technology at MCNC. He played a central role in establishing the NC Research and Education Network and the North Carolina Supercomputing Center. His education includes an MBA from the Fuqua School of Business, an MA in Liberal Studies from Duke University, and a ThM and an MDiv from Princeton Theological Seminary.

Darryl V. Creel, MS, is a Senior Research Statistician at RTI International. He has 19 years of experience working in the statistical field with a focus on survey sampling and biostatistical research. His general statistical experience includes survey sampling, data visualization, data analysis, and statistical learning. His statistical experience has been applied to a wide spectrum of research projects, ranging from large national surveys to relatively small surveys of highly specific populations.

Kathleen N. Lohr, PhD, MPhil, MA, is a Distinguished Fellow at RTI International with more than 40 years of experience in health care and health care policy research. She was the founding director of (and is now senior advisor to) the RTI–University of North Carolina (UNC) Evidence-based Practice Center. Dr. Lohr was also the founding editor-in-chief of RTI Press, a corporate-wide initiative. In 2007, RTI honored her with the Margaret Elliott Knox Excellence Award. Before coming to RTI, her 9 years at the Institute of

Medicine, National Academy of Sciences, entailed overseeing the Division of Health Care Services' portfolio of studies in health care delivery, organization, financing, quality of care and clinical evaluation, practice guidelines, health workforce, public health, and related topics. During her 12 years at the RAND Corporation, she led or participated in numerous health care projects for the US Department of Health and Human Services, Department of Defense, and the congressional Office of Technology Assessment. She is the author or coauthor or more than 350 books, monographs, chapters, journal articles, and other publications. In 2005, she was awarded the Avedis Donabedian Outcomes Research Lifetime Achievement Award by the International Society of Pharmacoeconomics and Outcomes Research. She served a 3-year term (2008–2010) as a member of the National Advisory Council, Agency for Healthcare Research and Quality, and she continues to serve on several expert panels and advisory boards for national and international groups. For several years at UNC-Chapel Hill, Dr. Lohr held the rank of Research Professor, Health Policy and Administration; she remains an adjunct professor in Health Policy and Management in the Gillings School of Global Public Health and a Senior Research Fellow at the UNC Cecil G. Sheps Center for Health Services Research. In 2016, she was appointed Adjunct Research Professor in the Department of Medical Informatics and Clinical Epidemiology at Oregon Health & Science University in Portland.

Edo D. Pellizzari, PhD, is Lead Fellow (Emeritus) in the RTI International Fellow Program. His nearly 40 years of research experience at RTI spans biochemistry, analytical chemistry, and environmental health. His research focused on exposure of children and adults to toxic chemicals such as volatile organic chemicals (e.g., benzene, chloroform, trichloroethylene, tetrachloroethylene, and bromodichloromethane), polynuclear aromatic hydrocarbons (e.g., benzopyrene), polychlorinated biphenyls, and heavy metals (e.g., lead, mercury, cadmium, arsenic, and manganese in air, food, and drinking water), and PM_{10} and $PM_{2.5}$ particles in air. Dr. Pellizzari developed qualitative and quantitative analytical chemistry methods and questionnaires for measuring personal exposure. He applied these methods to 14 probability-based human population studies throughout the United States and Canada. Dr. Pellizzari has served on several National Research Council committees; as chair of the Environmental Health Sciences Committee for the National Institute of Environmental Health Sciences (NIEHS) from 2006 to 2008; as a member of the Health Effects Institute (HEI) from 1998 to 2007; and as a

member of the US EPA Science Advisory Board's Drinking Water Committee from 1987 to 1997. He has served on 40 scientific review panels for NIEHS, EPA, the Department of Energy, and the National Institute of Occupational Safety and Health. He has authored or coauthored more than 240 scientific publications. He received his AB in biology from California State University, Chico, and his PhD in biochemistry from Purdue University; he was also a Fulbright-Hays recipient and a Fellow of the US Public Health Service.

Appendix A
Reproducible Research: Federally Regulated and Nonregulated Research

A significant proportion of research performed in the United States is not subject to federal guidelines or regulations. Notably, the reproducible research crisis is found predominantly in the nonregulated domain. Research conforming to federal regulatory requirements has standardized, long-standing definitions for terms such as *reproducibility, repeatability, replication,* and *accuracy* that aid in accomplishing reproducible research. Nonregulatory research, by contrast, has no comparable driving force to define a lexicon.

As discussed in this monograph, considerable discussion and confusion are occurring across numerous scientific disciplines about defining and achieving reproducible research in the nonregulated domain. Given this state of affairs, in this appendix we discuss federally regulated concepts and systems for quality research because they may be useful for consistently realizing reproducible research in the nonregulated arena.

Generally speaking, "regulated research" resides more in the physical and natural sciences than in the others. Thus, the points below pertain more to the natural and physical sciences (i.e., bench science) than to the social sciences (broadly defined). Many rules and guidelines pertain to the social sciences (here including clinical and social experiments as well as observational studies of all types)—especially when the studies involve human subjects. We discuss them briefly in this appendix (as "nonregulated research"). Nevertheless, in some cases, researchers in areas outside the bench sciences may well be conducting research projects for regulatory agencies and thus follow, to one degree or another, the precepts noted here.

Federally Regulated Research

Promulgated Definitions, Standards, and Systems

Federal guidelines apply when researchers are producing research results that are to be used for federal regulatory purposes. The US Food and Drug Administration (FDA) and the US Environmental Protection Agency (EPA) adhere to Good Laboratory Practice guidelines (GLPs) promulgated by the federal government. The FDA also adheres to Good Manufacturing Practices (GMPs). The EPA adheres to GLPs per the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) and the Toxic Substances Control Act (TSCA). It also follows EPA Quality System documents such as *EPA Requirements for Quality Management Plans (QA/R-2)* (publication number EPA/240/B-01/002) and *EPA Requirements for Quality Assurance Project Plans (QA/R-5)* (publication number EPA/240/B-01/003).

Research involving international applied research and analysis may also need to comply with various quality guidelines. These may come from, for example, the Organisation for Economic Co-operation and Development (OECD).

In addition, the following US agencies or associations have issued definitions and standards for research quality:

- the National Institute of Science and Technology (NIST)
- the American National Standards Institute (ANSI)
- the EPA, particularly quality management plans and quality assurance project plans
- the Society of Quality Assurance.

More generally, federal regulations—particularly the Federal Acquisition Regulations, which provides uniform policies and procedures for acquisitions by executive agencies of the federal government (48 CFR Part 46—Quality Assurance)—codify a wide array of standards.

Quality Assurance and Quality Control

Both quality assurance (QA) and quality control (QC) are cornerstone systems of federally regulated research and should be cornerstones of nonregulated research, but there is no standard. QA, generally speaking, is a systematic activity undertaken to ensure that research and services meet client-specific requirements. QC involves the operational aspects of a quality system that are

aimed at monitoring work processes and detecting and eliminating the causes of unsatisfactory performance as the research is being performed.

Many quality systems also use assessments, audits, and reviews of lessons learned. In keeping with the philosophy implied by the types of standards that agencies promulgate, we take the position (as in Chapter 4) that quality (QA/QC) should be built into the research conducted from the outset.

International groups issue similar guidance. They include the International Organization for Standardization, the Research Quality Association (RQA, formerly BARQA, the British Association of Research Quality Assurance), and the World Health Organization (WHO).

Other groups (domestic and international) have promulgated standards for GLPs and GMPs, as well as good clinical practices (GCPs). The Lean Six Sigma Institute has similar programs in both the United States and abroad.

Office of Quality Assurance

Federally regulated research requires that an independent Office of Quality Assurance provide oversight for implementation of controlled documents, such as standard operating procedures (SOPs), work practice documents, quality management plans, and quality assurance project plans. This office performs audits to ensure that research staff are following the instructions as prescribed for SOPs and use of notebooks and work practice document.

Enabling Reproducible Research Through Transparency

Given the varying definitions and the myriad issues raised throughout this monograph, describing how best to conduct either reproducible or at least transparent research is challenging. The core steps appear to lie in full and accurate documentation of the original work or, for that matter, "reproduced" or "replicated" studies. The central ideas regarding reproducibility, replicability, repeatability, accuracy, and validity apply across research (again broadly defined) in the social, natural, and physical sciences; for that reason, we have tried to make the ideas seamless across most fields and disciplines.

A tenet of regulated research is to document all steps employed in a study to provide complete transparency of how the work was performed and so that others can scrutinize and evaluate its validity and quality. Specifically, the concepts of reproducibility, replicability, repeatability, accuracy, and validity, as well as others, are to be proactively and qualitatively and/or quantitatively addressed during an experiment or study. Regulations explicitly standardize the definitions of these terms.

The fundamentals of regulatory research strive to achieve and display these constructs through the use of (1) SOPs, (2) notebooks to record what is done during each step in a study, (3) independent audits, and (4) measurement standards, where applicable. These fundamentals are noted here.

Standard Operating Procedures

SOPs should be consistent with appropriate regulations from the relevant regulatory agency. SOPs are written procedures, developed before any research commences, that document how an activity is implemented. When made available to the client and the public, they provide transparency that enables reproducible research. SOPs are prescriptive in nature; that is, they provide exacting detail regarding how study components are to be done. Normally, a description of how to implement a component of a study in an SOP is based upon prior knowledge or experience.

SOPs address the elements of reproducibility, replicability, repeatability, accuracy, validation (e.g., the accuracy and reproducibility of an instrument), and verification—essentially, all operational components of a study. QA or QC steps are included in the SOPs to assess the performance of instruments, devices, or systems, especially reproducibility, repeatability, and accuracy.

For example, in the natural and physical sciences, SOPs may be used for analytical methods, method development and analysis of biological samples, electrophoresis, system suitability, and column performance for liquid and gas chromatography, to name a few. Analytical laboratory techniques are described, including test material receipt, handling, and storage; labeling of reagents and solutions; formulation of drug substances; and water purification procedures, melting point determinations, and boiling point determinations. Data management SOPs include quality control of chemistry data.

Work for the commercial pharmaceutical industry has additional regulatory requirements. SOPs or work practice documents are employed and audits are performed to ensure that researchers are following the procedures as prescribed. These practices provide transparency and enable reproducible research.

Notebooks and Work Practice Documents

In addition to SOPs, researchers use laboratory or field (e.g., environmental work) notebooks or work practice documents in regulated research. Laboratory or field notebooks are essentially diaries; they capture in narrative form what and how a researcher conducts research from its inception to

its finality. The written narrative describes the objectives or hypothesis, experimental or study design, quality control, methods and equipment used to generate data, data analysis methods used, statistical algorithms used, and the interpretation of results. Each page of the notebook is signed and dated by a QA specialist in the organization's Office of Quality Assurance to affirm that the narrative clearly states what and how the work was done.

Often, investigators use both SOPs and notebooks in combination. Specifically, SOPs prescribe what steps to perform and how to perform them in research work, and notebooks document exactly what was done. In addition, notebooks let staff record any slight deviation from the SOP or complications that occurred while implementing the SOP. Often, SOPs are amended to improve them based on actual experience from implementing them. Use of SOPs and laboratory or field notebooks enables the research to be reproduced.

Use of notebooks is mandatory when conducting regulated research. This habit is engrained in students during their graduate training, and it continues when conducting regulated and, in some cases, nonregulated research.

Audits

In the natural and physical sciences, quality assurance unit staff conduct internal audits and FDA and EPA staff conduct external audits to ensure that detailed SOPs exist for all regulated studies. Furthermore, audits are done to determine whether investigators are following the SOPs as prescribed and to evaluate the performance of methods and instruments for replication, reproducibility, and accuracy. Audits provide independent verification that instruments are performing accurately against measurement standards from NIST, clients, or commercial repositories.

Nonregulated Research

Nonregulated research generally does not conform to any detailed federal standards of the sort described above regarding how research is to be documented. Nevertheless, numerous agencies have expectations about protocols or specific rules concerning (for instance) informed consent, human subject protections, approval by the US Office of Management and Budget for questionnaires and other paperwork, or conduct of systematic reviews of the published literature. Scientists meet these expectations by documenting the details of their research in notebooks or work practice documents—in some cases perhaps less comprehensively for research-sponsoring organizations per

se than for the peer-reviewed journals in which staff may want to report their work.

Some organizations that support nonregulated or basic research projects may also require QA systems or activities. For example, the US Agency for Healthcare Research and Quality (AHRQ) asks its evidence-based practice centers specifically how they will exercise what amounts to QA efforts for producing systematic reviews and technical briefs.

The concepts of QA, QC, and quality planning are practiced intermittently across domestic and international nonregulated research. QA elements may include training on quality and documentation during the start-up of projects. QC involves monitoring and evaluation activities and audits to ensure that research is well documented. Thus, these practices provide transparency as well, which enables reproducible research and analysis.

Protocols

As for SOPs, investigators in nonregulated research programs increasingly create and follow written protocols. Protocols may contain many of the characteristics of SOPs. They describe the approach investigators will use to perform the research in question. However, they are not specified under the umbrella of a regulatory requirement. Protocols may address elements of reproducibility, repeatability, replication, and accuracy. By their very nature, protocols also provide transparency to how work is performed and thus can enable research to be reproduced. Organizations and agencies that require protocols (and perhaps registration of those protocols in a US or international registry) include the US Patient-Centered Outcomes Research Institute (PCORI), AHRQ, and the US National Institutes of Health (NIH).

Standard Operating Procedures

Although formal SOPs may be used more rarely in nonregulated research, they still can have a role across all kinds of investigations. Professional societies in many disciplines may have recommendations concerning such practices for certain applications. For example, SOPs apply to survey work at RTI International (e.g., "Deterrence of Falsification by Data Collectors," "Detection of Falsification by Data Collectors," and "Remediation of Falsification by Data Collectors"). Authoritative national bodies may also issue reports defining standards for good practices (as from the Health and Medicine Division of the National Academies of Sciences, Engineering, and Medicine). The PCORI Methodology Committee has issued a methodology standards report that

lays out expectations for the methods to be followed in conducting patient-centered research in the health arena. The American Association for Public Opinion Research has similar requirements.

Notebooks

In nonregulated research, laboratory and field notebooks are more often used than SOPs. "Notebooks," as a concept for the social sciences, may have a broader connotation—namely, documentation done in Microsoft Excel or Word files or in coding manuals. For literature or systematic reviews, for example, detailed documentation from searches in MEDLINE are typically required (search terms, order of application, and yields).

Using notebooks (in this broader sense) is left to the discretion of the project's principal investigator except insofar as agencies may require documentation (especially in contract work). As above, investigators may well specify protocols and then use both SOPs and notebooks in combination.

Other Tools

Many domains of science have several options for software tools to help overcome the technical challenges of doing reproducible research.[81] These tools enable researchers to capture and communicate the details of their workflow with much greater efficiency than simply writing a lengthy prose narrative. Thus, these tools can provide transparency that may permit achieving reproducible research. Numerous other software tools have been developed for documenting computational work. They may be grouped into five general types: literate computing, authoring, and publishing; controlling versions of documents; tracking provenance; automating steps; and capturing the computational environment.

Increased reliance on computational approaches in the life sciences has revealed grave concerns about how accessible and reproducible computation-reliant results truly are. To this end, a specific example tool comes from Pennsylvania State University's Center for Comparative Genomics and Bioinformatics and Johns Hopkins University's Department of Biology. Together they developed Galaxy, a scientific workflow system to support accessible, reproducible, and transparent computational research in the life sciences. Specifically, Galaxy (https://usegalaxy.org) is an open web-based platform for genomic research.[124] Galaxy automatically tracks and manages data provenance and provides support for capturing the context and intent of

computational methods. Galaxy Pages provide users with a medium by which to communicate a complete computational analysis.

Documenting Reproducible Research

Reproducible research entails creating and providing access to complete descriptions of the study designs, critical assumptions, methods, tools (e.g., for data collection), source code or specifics of statistical techniques employed, and the resulting data of the original work. Such documentation must be sufficient to enable other investigators, independently, to attempt to reproduce the initially reported results. This expectation may, however, be constrained by factors such as permissions granted by institutional review boards (IRBs), data use agreements (e.g., with federal agencies or commercial clients), or other client-related prohibitions.

To fulfill this expectation about documentation, researchers can create a package of requisite materials that describes the research process and the results. Such a package becomes a complete guide for both reproducing and extending the research. In some cases, much of the descriptive material (excluding results) may be found in research protocols registered with relevant public-sector agencies (e.g., https://clinicaltrials.gov) maintained by NIH.

Private-sector entities also provide such services. For example, the Centre for Reviews and Dissemination at the University of York in the UK created PROSPERO (http://www.crd.york.ac.uk/PROSPERO), an international registry for systematic reviews. It captures features from review protocols in health and social care, welfare, public health, education, crime, justice, and international development that have some form of a health-related or patient-centered outcome. In addition, descriptive materials documenting at least some steps may be found as supplemental web-based information (e.g., appendices) to papers published in peer-reviewed journals or various other authoritative reports.

Appendix B
Transparent Research

Transparency in Research: Understanding Basic Categories

The categories introduced here relate to reproducible—or more accurately, transparent—research and cover many typical research contexts. Adapted from Stodden et al.[76] and Brandt et al.,[38] they pertain to a wide range of research in the social, natural, and physical sciences.

Reviewable research. Descriptions of the research methods can be independently assessed and the results judged credible (or not). This formulation of "reviewable" includes both traditional peer review and community review. This concept does not necessarily imply reproducibility.

Replicable research. Investigators make available tools to allow other researchers to duplicate the results of their research, using precisely the same methods and procedures throughout. One example might be running the authors' code to produce the plots shown in any publication. This particular process could be long and arduous, however; for that reason, investigators rarely take this approach. A back-up step is for investigators to make sure that they have adequately documented all their codes.

Confirmable research. The main interpretations of the research can be attained independently without the use of any software or other tools that research teams (or authors) might produce. For example, other investigators can reproduce results by using the complete description of algorithms and methodology that the original researchers provide in any publications or supplementary materials.

Auditable research. Researchers have archived sufficient records (including data and software) so that they can defend their research later if necessary or help to resolve differences when repeated by independent researchers.

Researchers may face special circumstances related to auditing that they can solve in different ways.

For example, the archive might be private. This is the case, for instance, with traditional laboratory notebooks, which may need to be made available upon request from US regulatory agencies such as the Environmental Protection Agency (EPA), the Food and Drug Administration (FDA), or the National Toxicology Program (NTP) of the National Institute of Environmental Health Sciences. In these circumstances, EPA, FDA, or NTP auditors may come to the researchers' laboratories or institutions to audit the notebooks. Investigators might also place such information in a more public archive such as a repository or registry, as noted in Appendix A.

Open or reproducible research. Researchers make auditable research openly available. This category comprises well-documented and fully open code and data that are publicly available and that would allow others to

- fully audit the study procedures and computations
- replicate or also independently reproduce the results of the research
- extend the results or apply the method to new problems.

Verification. This category entails checking that the computer code that the original investigators describe correctly solves the mathematical problem it claims to solve. Those verifying such code also need to examine whether the code is doing what it is supposed to be doing. For example, verification could be something as simple as checking that a variable was recoded as described in the text or as complex as a modeling function or procedure having the correct options set as described in the text.

For standardized software, e.g., SAS, Stata, or other commercial programs, validating the underlying algorithms that produce the numerical results may not be needed or appropriate. Instead, checking to ensure that the options that were supposed to be used were actually used would be appropriate. For custom software, verification requires examining whether algorithms are functioning correctly.

Validation. This category includes checking that the results of a computer simulation agree with experiments or observations of the phenomenon being studied. This element applies only if the model represents an observable event. Many investigators may, however, try to model events that have not yet happened or that occurred many years in the past. These situations pose

difficulties that those using validation measures need to acknowledge and attempt to address.

Transparency in Research: Identifying Limitations

This spectrum of research transparency allows both investigators and end users of research to better understand current practices for conducting a full range of natural, physical, and social science research. These categories outline what scientists can define and describe about their research. Such classes of activities help to identify limitations in any or all of the following eight components of high-quality research:

- assumptions made as part of designing a study
- methods, procedures, and tools used to generate data
- how missing data are handled
- how data are curated (i.e., managed, annotated, reported, and kept safe and available for reuse)
- how restricted-use, protected, or personally identifiable information is used while protecting confidentiality and privacy
- analytic tools used to perform analyses
- quality control methods used in each step of the studies
- interpretations of the results and conclusions drawn from those interpretations.

Identifying such limitations points to ways that researchers can make basic changes in their work to improve reproducibility. Such transparency helps to forestall reasons that investigators might use not to release replication or reproducibility "packages."

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452-454. https://doi.org/10.1038/533452a
2. Popper KR. The logic of scientific discovery. London, UK: Hutchinson; 1959.
3. Journals unite for reproducibility. Nature. 2014;515:7. https://doi.org/10.1038/515007a
4. Van Noorden R. Parasite test shows where validation studies can go wrong. Nature. 2014. https://doi.org/10.1038/nature.2014.16527
5. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014;505(7485):612-3. https://doi.org/10.1038/505612a
6. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012;483(7391):531-3. https://doi.org/10.1038/483531a
7. Laman JD, Kooistra SM, Clausen BE. Reproducibility issues: avoiding pitfalls in animal inflammation models. Methods Mol Biol. 2017;1559:1-17. https://doi.org/10.1007/978-1-4939-6786-5_1
8. Easley RW, Madden CS. Replication revisited: introduction to the special section on replication in business research. J Business Res. 2013;66(9):1375-1376. https://doi.org/10.1016/j.jbusres.2012.05.001
9. Easley RW, Madden CS, Gray V. A tale of two cultures: revisiting journal editors' views of replication research. J Business Res. 2013;66(9):1457-1459. https://doi.org/10.1016/j.jbusres.2012.05.013
10. Evanschitzky H, Armstrong JS. Research with in-built replications: comment and further suggestions for replication research. J Business Res. 2013;66(9):1406-1408. https://doi.org/10.1016/j.jbusres.2012.05.006
11. Uncles MD, Kwok S. Reply to commentary on designing research with in-built differentiated replication. J Business Res. 2013;66(9):1409-1410. https://doi.org/10.1016/j.jbusres.2012.05.007
12. Uncles MD, Kwok S. Designing research with in-built differentiated replication. J Business Res. 2013;66(9):1398-1405. https://doi.org/10.1016/j.jbusres.2012.05.005
13. LeVeque RJ, Mitchell IM, Stodden V. Reproducible research for scientific computing: tools and strategies for changing the culture. Comput Sci Eng. 2012;14(4):13-17. https://doi.org/10.1109/MCSE.2012.38
14. Donoho DL. An invitation to reproducible computational research. Biostatistics. 2010;11(3):385-8. https://doi.org/10.1093/biostatistics/kxq028

15. Mullard A. Reliability of 'new drug target' claims called into question. Nat Rev Drug Discov. 2011;10(9):643-644. https://doi.org/10.1038/nrd3545

16. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016;351(6280):1433-6. https://doi.org/10.1126/science.aaf0918

17. Chang AC, Li P. Is economics research replicable? Sixty published papers from thirteen journals say "usually not." Washington, DC: Board of Governors of the Federal Reserve System; 2015. Finance Econ Discuss Ser 2015-083. https://doi.org/10.17016/feds.2015.083

18. Cook BG. A call for examining replication and bias in special education research. Remedial Special Educ. 2014;35(4):233-246. https://doi.org/10.1177/0741932514528995

19. Makel MC, Plucker JA. An introduction to replication research in gifted education: shiny and new is not the same as useful. Gifted Child Q. 2015;59(3):157-164. https://doi.org/10.1177/0016986215578747

20. Makel MC, Plucker JA. Facts are more important than novelty: replication in the education sciences. Educ Res. 2014;43(6):304-316. https://doi.org/10.3102/0013189x14545513

21. Spector JM, Johnson TE, Young PA. An editorial on replication studies and scaling up efforts. Educ Tech Res Dev. 2015;63(1):1-4. https://doi.org/10.1007/s11423-014-9364-3

22. Warne RT. Two additional suggested reforms to encourage replication studies in educational research. Educ Res. 2014;43(9):465-465. https://doi.org/10.3102/0013189x14562294

23. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. Am J Epidemiol. 2006;163(9):783-9. https://doi.org/10.1093/aje/kwj093

24. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Pooles C, Goodman SN, et al. Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337. https://doi.org/10.1007/s10654-016-0149-3

25. Boos DD, Stefanski LA. P-value precision and reproducibility. Am Stat. 2011;65(4):213-221. https://doi.org/10.1198/tas.2011.10129

26. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype-phenotype associations. Nature. 2007;447(7145):655-60. https://doi.org/10.1038/447655a

27. Casadevall A, Fang FC. Reproducible science. Infect Immun. 2010;78(12):4972-5. https://doi.org/10.1128/IAI.00908-10

28. Morrell K, Lucas JW. The replication problem and its implications for policy studies. Crit Policy Stud. 2012;6(2):182-200. https://doi.org/10.1080/19460171.2012.689738

29. Ishiyama J. Replication, research transparency, and journal publications: individualism, community models, and the future of replication studies. PS Polit Sci Polit. 2013;47(01):78-83. https://doi.org/10.1017/s1049096513001765

30. Carsey TM. Making DA-RT a reality. PS Polit Sci Polit. 2013;47(01):72-77. https://doi.org/10.1017/s1049096513001753

31. Dafoe A. Science deserves better: the imperative to share complete replication files. PS Polit Sci Polit. 2013;47(01):60-66. https://doi.org/10.1017/s104909651300173x

32. Elman C, Kapiszewski D. Data access and research transparency in the qualitative tradition. PS Polit Sci Polit. 2013;47(01):43-47. https://doi.org/10.1017/s1049096513001777

33. Lupia A, Alter G. Data access and research transparency in the quantitative tradition. PS Polit Sci Polit. 2013;47(01):54-59. https://doi.org/10.1017/s1049096513001728

34. McDermott R. Research transparency and data archiving for experiments. PS Polit Sci Polit. 2013;47(01):67-71. https://doi.org/10.1017/s1049096513001741

35. Moravcsik A. Transparency: the revolution in qualitative research. PS Polit Sci Polit. 2013;47(01):48-53. https://doi.org/10.1017/s1049096513001789

36. Lupia A, Elman C. Openness in political science: data access and research transparency. PS Polit Sci Polit. 2013;47(01):19-42. https://doi.org/10.1017/s1049096513001716

37. Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, Fiedler K, et al. Recommendations for increasing replicability in psychology. Eur J Personality. 2013;27(2):108-119. https://doi.org/10.1002/per.1919

38. Bonett DG. Replication-extension studies. Curr Dir Psychol Sci. 2012;21(6):409-412. https://doi.org/10.1177/0963721412459512

39. Brandt MJ, Ijzerman H, Dijksterhuis A, Farach FJ, Geller J, Giner-Sorolla R, et al. The replication recipe: what makes for a convincing replication? J Exp Soc Psychol. 2014;50:217-224. https://doi.org/10.1016/j.jesp.2013.10.005

40. Koole SL, Lakens D. Rewarding replications: a sure and simple way to improve psychological science. Perspect Psychol Sci. 2012;7(6):608-14. https://doi.org/10.1177/1745691612462586

41. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716. https://doi.org/10.1126/science.aac4716

42. Kappenman ES, Keil A. Introduction to the special issue on recentering science: replication, robustness, and reproducibility in psychophysiology. Psychophysiology. 2017;54(1):3-5. https://doi.org/10.1111/psyp.12787

43. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafo MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci. 2017;18(2):115-126. https://doi.org/10.1038/nrn.2016.167

44. Lucas J, W., Morrell K, Posard M. Considerations on the "replication problem" in sociology. American Sociologist. 2013;44(2):217-232. https://doi.org/10.1007/s12108-013-9176

45. Baker M. Over half of psychology studies fail reproducibility test. Nature News. 2015 Aug 27. https://doi.org/10.1038/nature.2015.18248

46. Baker M. Psychology's reproducibility problem is exaggerated – say psychologists. Nature. 2016. https://doi.org/10.1038/nature.2016.19498

47. Gilbert DT, King G, Pettigrew S, Wilson TD. Comment on "Estimating the reproducibility of psychological science." Science. 2016;351(6277):1037. https://doi.org/10.1126/science.aad7243

48. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:0021. https://doi.org/10.1038/s41562-016-0021

49. Resnik DB, Shamoo AE. Reproducibility and research integrity. Accountability Res. 2017;24(2):116-123. https://doi.org/10.1080/08989621.2016.1257387

50. Donoho DL, Maleki A, Rahman IU, Shahram M, Stodden V. Reproducible research in computational harmonic analysis. Comput Sci Eng. 2009;11(1):8-18. https://doi.org/10.1109/MCSE.2009.15

51. Holdren J. Increasing access to the results of federally funding scientific research. Memo to the heads of executive departments and agencies. [2013 Feb 22; cited 2017 Feb 20]; Available from: https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

52. National Science Foundation. Today's data, tomorrow's discoveries: increasing access to the results of research funded by the National Science Foundation. NSF public access plan (NSF 15-52). [2015 Mar 18; cited 2016 Aug 22]; Available from: https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf

53. Buck S. Comments in response to "Request for information re: strategy for American innovation." [2014 Sep 23; cited 2017 Apr 3]; Available from: http://www.arnoldfoundation.org/wp-content/uploads/2015/05/Comments-on-Scientific-Reproducibility1.pdf

54. Peng R. The reproducibility crisis in science: a statistical counterattack. Significance. 2015;12(3):30-32. https://doi.org/10.1111/j.1740-9713.2015.00827.x

55. Herndon T, Ash M, Pollin R. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge J Econ. 2013;38(2):257-279. https://doi.org/10.1093/cje/bet075

56. Cheng S, Powell B. Measurement, methods, and divergent patterns: reassessing the effects of same-sex parents. Soc Sci Res. 2015;52:615-26. https://doi.org/10.1016/j.ssresearch.2015.04.005

57. Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009;41(2):149-55. https://doi.org/10.1038/ng.295

58. McNutt M. Editorial expression of concern. Science. 2015;348(6239):1100. https://doi.org/10.1126/science.aac6184

59. Nuzzo R. How scientists fool themselves—and how they can stop. Nature. 2015;526(7572):182-5. https://doi.org/10.1038/526182a

60. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293-314. https://doi.org/10.1093/nsr/nwt032

61. Temchine M. Reproducibility will not cure what ails science. Nature. 2015;525(159):1-8. https://doi.org/10.1038/525159a

62. National Academies of Sciences, Engineering, and Medicine. Fostering integrity in research. Washington, DC: National Academies Press; 2017. https://doi.org/10.17226/21896

63. Carley KM, Frantz TL, eds. Computational and mathematical organization theory. [cited 2017 Mar 28]; Available from: https://link.springer.com/journal/10588

64. McNutt M. Journals unite for reproducibility. Science. 2014;346(6210):679-679. https://doi.org/10.1126/science.aaa1724

65. Karr AF. Why data availability is such a hard problem. Stat J IAOS. 2014;30:101-107.

66. Baker M. Muddled meanings hamper efforts to fix reproducibility crisis. Nature. 2016. https://doi.org/10.1038/nature.2016.20076

67. Bissell M. Reproducibility: the risks of the replication drive. Nature. 2013;503(7476):333-4. https://doi.org/10.1038/503333a

68. National Institute of Standards and Technology. NIST Information Quality Standards. c2009. [2016 Sep 1; cited 2017 Jan 4]; Available from: https://www.nist.gov/nist-information-quality-standards

69. English Oxford Living Dictionaries [Internet]. Oxford, UK: Oxford University Press; 2017. Available from: https://en.oxforddictionaries.com/

70. Taylor BN, Kuyatt CE. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST Technical Note 1297. 1994 ed. Gaithersburg, MD: National Institute of Standards and Technology.

71. Joint Committee for Guides in Metrology (JCGM 200:2012). International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 3rd ed. 2008 version with minor corrections. Sèvres, France: Bureau International des Poids et Mesures (BIPM); 2012. Available from: http://www.bipm.org/en/publications/guides/vim.html

72. MiC Quality. Six Sigma glossary. Repeatability & reproducibility [online]. [No date; cited 2017 Mar 13]; Available from: http://www.micquality.com/six_sigma_glossary/repeatability_reproducibility.htm

73. Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL. Social, behavioral, and economic sciences perspectives on robust and reliable science. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. [2015 May; cited 2017 Jan 4]; Available from: https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

74. Leek JT, Peng RD. Opinion: reproducible research can still be wrong: adopting a prevention approach. Proc Nat Acad Sci. 2015;112(6):1645-1646. https://doi.org/10.1073/pnas.1421412111

75. McNaught AD, Wilkinson A, compilers. International Union of Pure and Applied Chemistry. Compendium of chemical terminology. 2nd ed. Oxford, UK: Blackwell Science; 1997.

76. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11:193-205.

77. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

78. Stodden V. 2014: what scientific idea is ready for retirement? Reproducibility. [2015; cited 2017 May 24]; Available from: http://edge.org/response-detail/25340

79. RTI International. Quality manual: a corporate overview of quality. Version 4.0. Research Triangle Park, NC: RTI International; 2016.

80. Institute of Medicine (US) Committee to Design a Strategy for Quality Review and Assurance in Medicare; Lohr KN, editor. Medicare: a strategy for quality assurance. Vol. I. Washington, DC: National Academies Press; 1990.

81. Knox L, Brach C. The practice facilitation handbook: training modules for new facilitators and their trainers. Rockville, MD: Agency for Healthcare Research & Quality; 2013. Available from: https://www.ahrq.gov/sites/default/files/publications/files/practicefacilitationhandbook.pdf

82. Stodden V, Bailey DH, Borwein J, LeVeque RJ, Rider W, Stein W. Setting the default to reproducible: reproducibility in computational and experimental mathematics. Providence, RI: ICERM Workshop Report; 2013. Available from: https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf

83. Cella D, Hahn E, Jensen S, Butt Z, Nowinski C, Rothrock N, et al. Patient-reported outcomes in performance measurement (RTI Press Publication No. BK-0014-1509). Research Triangle Park, NC: RTI Press; 2015. https://doi.org/10.3768/rtipress.2015.bk.0014.1509

84. Cochrane Collaboration. Cochrane Collaboration glossary version 4.2.5. [2005 May; cited 2011 January]; Available from: http://www.cochrane.org/sites/default/files/uploads/glossary.pdf; http://effectivehealthcare.ahrq.gov/

85. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Davey-Smith SG, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. 2nd ed. London, United Kingdom: BMJ Books; 2001. p. 87-108.

86. Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters LM, et al. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. Methods guide for comparative effectiveness reviews. AHRQ Publication No. 12-EHC047-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2012.

87. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care. 2004;16(1):9-18. https://doi.org/10.1093/intqhc/mzh005

88. Shadish W, Cook T, Campbell D. Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin; 2002.

89. Atkins D, Chang S, Gartlehner G, Buckley DI, Whitlock EP, Berliner E, et al. Assessing the applicability of studies when comparing medical interventions. Methods guide for effectiveness and comparative effectiveness reviews. AHRQ publication No. 11-EHC019-EF. [2010 Dec 30; cited 2017 Mar 28]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK53480/

90. Donabedian A. Evaluating the quality of medical care. Milbank Mem Fund Q. 1966;44(3):Suppl:166-206.

91. Donabedian A. Explorations in quality assessment and monitoring. Vol. 1. The definition of quality and approaches to its assessment. Ann Arbor, MI: Health Administration Press; 1980.

92. Donabedian A. Explorations in quality assessment and monitoring. Vol. II. The criteria and standards of quality.Ann Arbor, MI: Health Administration Press; 1982.

93. Donabedian A. Explorations in quality assessment and monitoring. Vol. III. The methods and findings of quality assessment and monitoring: an illustrated analysis. Ann Arbor, MI: Health Administration Press; 1984.

94. Deming WE. Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology Press; 1986.

95. Garvin DA. A note on quality: the views of Deming, Juran, and Crosby. Harvard Business School Note 9-687-011. Cambridge, MA: Harvard Business School; 1986.

96. Walton M. The Deming management method. New York, NY: Dodd, Mead; 1986.

97. Juran JM, Gyrna FM, Jr., Bingham RS, Jr. Quality control handbook. 4th ed. Manchester, MO: McGraw-Hill; 1988.

98. Berwick DM. Continuous improvement as an ideal in health care. New Engl J Med. 1989;320(1):53-56. https://doi.org/10.1056/Nejm198901053200110

99. Chassin MR. Is health care ready for Six Sigma quality? Milbank Q. 1998;76(4):565-91, 510. https://doi.org/10.1111/1468-0009.00106

100. Harry M, Schroeder R. Six Sigma. New York, NY: Doubleday Publishers; 2000.

101. DelliFraine JL, Langabeer JR, 2nd, Nembhard IM. Assessing the evidence of Six Sigma and Lean in the health care industry. Qual Manag Health Care. 2010;19(3):211-25. https://doi.org/10.1097/QMH.0b013e3181eb140e

102. Stodden V, Leisch F, Peng R, editors. Implementing reproducible research. Boca Raton, FL: CRC Plress; 2014.

103. Sackett DL, Straus SE, Richardson SR, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. London, UK: Churchill Livingstone; 1997.

104. Buckley DI, Ansari M, Butler M, Williams C, Chang C. The refinement of topics for systematic reviews: lessons and recommendations from the Effective Health Care Program. AHRQ Methods for Effective Health Care. [2013 Jan; cited 2017 Mar 28]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK121274/

105. Thompson M, Tiwari A, Fu R, Moe E, Buckley DI. A framework to facilitate the use of systematic reviews and meta-analyses in the design of primary research studies. AHRQ Methods for Effective Health Care. [2012 Jan; cited 2017 Mar 28]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK83621/

106. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews. 2015;4(1):1. https://doi.org/10.1186/2046-4053-4-1

107. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015;349:g7647. https://doi.org/10.1136/bmj.g7647

108. Drummond C. Replicability is not reproducibility: nor is it good science. Proc Eval Methods Mach Learn Workshop 26th ICML [International Conference on Machine Learning], Montreal, Quebec, Canada; 2009. Available from: http://cogprints.org/7691/7/icmlws09.pdf

109. Stodden V. Reproducing statistical results. Ann Rev Stat Apps. 2015;2:1-19. https://doi.org/10.1146/annurev-statistics-010814-020127

110. FitzJohn R, Pennell M, Zanne A, Cornwell W. Reproducible research is still a challenge. rOpenSci. 2014. Available from: http://ropensci.org/blog/2014/06/09/reproducibility/

111. Lohr KN. Scientific writing: making it readable. [2016; cited 2017 Jan16]; Available from: www.lohrconsulting.com

112. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187-91. https://doi.org/10.1038/nature11556

113. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010;8(6):e1000412. https://doi.org/10.1371/journal.pbio.1000412

114. Schulz KF, Altman DG, Moher D, Consort Group. CONSORT 2010 statement: guidelines for reporting parallel group randomised trials. J Clin Epidemiol. 2010;63(8):834-40. https://doi.org/10.1016/j.jclinepi.2010.02.005

115. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA. 2000;283(15):2008-2012. https://doi.org/10.1001/jama.283.15.2008

116. Moher D, Liberati A, Tetzlaff J, Altman DG, Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009;62(10):1006-12. https://doi.org/10.1016/j.jclinepi.2009.06.005

117. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol. 2009;62(10):e1-34. https://doi.org/10.1016/j.jclinepi.2009.06.006

118. Sackett DL. Superiority trials, noninferiority trials, and prisoners of the 2-sided null hypothesis. ACP J Club. 2004;140(2):A11-A12. https://doi.org/10.7326/ACPJC-2004-140-2-A11

119. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Ring D, et al. GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol. 2011;64(12):1283-1293. https://doi.org/10.1016/j.jclinepi.2011.01.012

120. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263(10):1385-9.

121. Berkman ND, Lohr KN, Ansari MT, Balk EM, Kane R, McDonagh M, et al. Grading the strength of a body of evidence when assessing health care interventions: an EPC update. J Clin Epidemiol. 2015;68(11):1312-24. https://doi.org/10.1016/j.jclinepi.2014.11.023

122. Kavvoura FK, Liberopoulos G, Ioannidis JP. Selection in reported epidemiological risks: an empirical assessment. PLoS Med. 2007;4(3):e79. https://doi.org/10.1371/journal.pmed.0040079

123. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010;340:c365. https://doi.org/10.1136/bmj.c365

124. Goecks J, Nekrutenko A, Taylor J, Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86. https://doi.org/10.1186/gb-2010-11-8-r86

Praise for

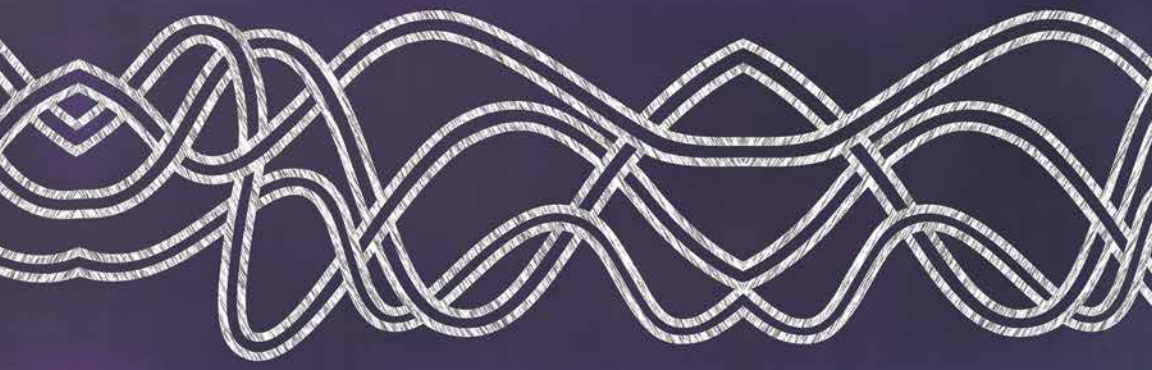
Reproducibility: A Primer on Semantics and Implications for Research

"The authors have written a nuanced and thoughtful primer on scientific reproducibility. By highlighting the social, political, and technical importance of reproducibility, together with a precise description of the related concepts of reproducibility, replicability, and repeatability, this primer provides a significant resource that all practicing researchers should read."

Daniel Reed, Vice President for Research and Economic Development, University of Iowa and former Corporate Vice President, Microsoft

"This is a well-written, clearly articulated, and timely primer on the developing and evolving rich terminology of reproducible research. The primer, put together by authors with deep experience and expertise in the topic area, focuses primarily on human-centric research in biomedicine, medicine, and the social sciences as well as reproducibility issues in analytics and computational science. The growing focus on reproducibility will open new vistas in research methodologies, meta analysis, comparative studies of research results, and reuse and adaptation of results from prior research. This primer provides an excellent overview of the subject area, and I would recommend it to anyone interested in coming up to speed on current issues in reproducible research."

Chaitan Baru, Distinguished Scientist and Associate Director for Data Initiatives, San Diego Supercomputing Center; current appointment as Senior Advisor for Data Science, Computer and Information Science and Engineering Directorate, National Science Foundation



"Pellizzari et al. have taken on the Herculean task of collecting, synthesizing, and relating the various interpretations of reproducibility used in the research community today, and turned the result into an accessible must-read guide. This important work provides a Rosetta Stone for various stakeholders to discuss and implement solutions that make real progress toward a research enterprise that routinely produces reproducible findings."

Victoria Stodden, Associate Professor at the School of Information Sciences, University of Illinois at Urbana Champaign and co-editor of the books *Implementing Reproducible Research* and *Privacy, Big Data, and the Public Good: Frameworks for Engagement*

RTI Press

RTI INTERNATIONAL

www.rti.org/rtipress

RTI Press publication BK-0020-1708