Replicability in Clinical Research

NBER-IFS International Network on the Value of Medical Research White Paper

Katherine Lofgren
Harvard University
Interfaculty Initiative in Health Policy

November 13, 2018

Abstract

Across scientific fields of inquiry, there is interest in the role of replication and confirmatory studies within the broader set of potential research objectives. The lack of replicability is of particular importance in pre-clinical and clinical research where research studies are expensive and study results inform both regulatory decisions and patient treatment. This white paper is intended as a resource to understand what we mean when we talk about replication. There are many types of actions that fall under the umbrella of replicability. Understanding what potential work may constitute replication and how to talk about the scope of replicability is a important communication tool. I also review strategies to define successful replication – even when the scope of work is well defined, there is little consensus about how to estimate the success of failure of a replication. The white paper also aims to summarize what we know about pre-clinical and clinical replicability. Particularly, the white paper discusses drug regulation as an empiric context to study the replicability of results through confirmatory trials. Although regulatory requirements for confirmatory trials are rarely framed as replications, they are just that — a space where decision making is conditional on successful or sufficient replicability of results.

Research reported in this publication was supported by the National Institute on Aging under Award Number R24AG048059 to the National Bureau of Economic Research. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health or NBER.

Table of Contents

Introduction

Academics, funding agencies, and the general public all care about the replicability of study results. It matters what we spend our money and time on. Ideally, effort and funding are allocated to the generation of results which improve decision making and reveal underlying truths. When we fall short of that goal, the repercussions can be dramatic – more research is called into question and, particularly in clinical settings, lives can be lost when ineffective or harmful therapies are used to treat patients.

Clinical research may intuitively seem like a relatively safe research space from replicability issues – much of the research is conducted as randomized controlled trials – the gold standard in research practice. However, pre-clinical and clinical research suffers from the same replicability issues observed across fields as varied as experimental psychology, economics, and political science. In two highly publicized articles, the pharmaceutical firms Bayer and Amgen both reported an inability to replicate the vast majority (75% and 89% failures respectively) of sampled study results (1,2). Clinical research is an expensive and time-consuming endeavor. The approval process for a new drug takes on average 12 years (3). Pre-clinical work that is erroneously promising, triggers further investments and commitment which ultimately may not produce approved, effective therapies. One indication of replicability issues is the failure rate of downstream clinical research. By some estimates, the probability of success in Phase II clinical trials is decreasing over time (4). Each step in the pre-clinical and clinical process increases the cost and time commitment of the study, when the replicability of early stage research is low, we pursue clinical research which does not produce benefits for society.

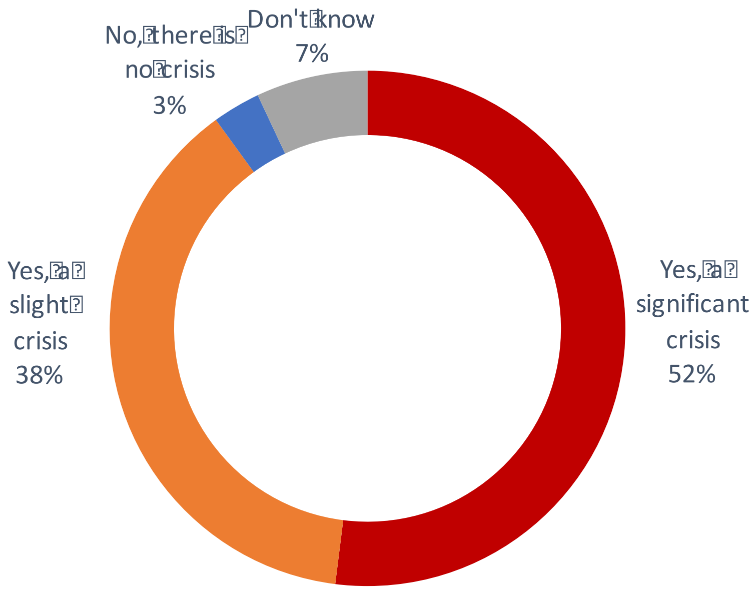
Beyond the low probability of success in the drug development pipeline, an even more troubling issue in clinical research is the use of ineffective (or harmful) therapies based on promising results during the regulatory process. One well-known example is the 2011 decision by the FDA to revoke the breast cancer indication for Avastin (5). Avastin was initially approved in 2008 through the FDA's accelerated approval pathway for the treatment of metastatic breast cancer in combination with paclitaxel for patients with HER2-negative breast cancer and no previous chemotherapy. In the year before the indication was revoked, 2010, there were 6.8 billion USD in sales recorded for the drug (5). In the case of Avastin, pre-approval studies measured the objective response rate in small patient samples (85 and 56 patients in two separate studies). When a confirmatory trial of a much larger study population, 921 patients, measured the true outcome of interest – overall survival – the implications of the earlier studies were not confirmed. Clinical research, regulatory, and treatment decisions necessarily occur under conditions of uncertainty. It's important to consider how to react to existing research. Do we feel comfortable given the information we currently have from the cumulative evidence available to:

(1) Wait, study the context more
(2) Act now, verify with additional formal study
(3) Act now, trust and allow observational information to accumulate

The choice to approve Avastin for the treatment of breast cancer on outcomes from two early studies while requiring a follow-up trial to measure overall survival falls under option 2. Had the early information been correct, patients would have had access to an effective treatment faster. Instead, patients were harmed from decision making based on surrogate endpoint outcomes. Weighing the balance between these potential futures (one where you were right all along vs. one where you were wrong) requires context specific considerations which quantify the upsides and downsides of incorrect decision making.

This white paper is intended as a general resource to better understand the academic conversation around replicability in clinical research. **Figure 1** summarizes the results of a survey conducted by Nature (2016) of 1,576 researchers across a variety of fields including biology and medicine (6). Ninety percent of respondents felt that there was either a slight or significant reproducibility crisis.

Figure 1: Nature (2016) Survey Results – "Is there a reproducibility crisis"?



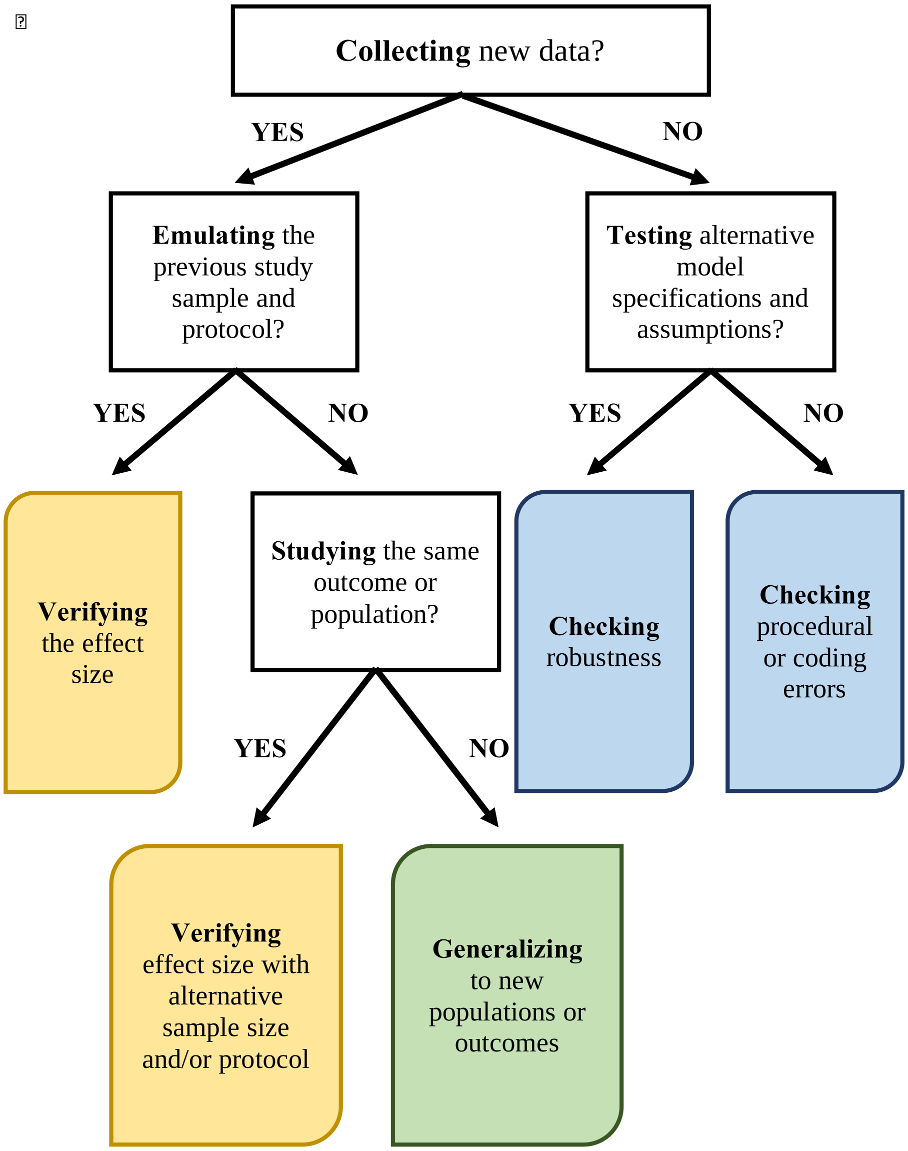
Despite a clear consensus that replicability and reproducibility are important and often lacking in research studies, there remains incredible ambiguity about what types of replication we need, when replication is worth conducting, and how to approach the analysis of replication efforts. This white paper is divided into three substantive sections: **Section 1** focuses on a review of definitions and types of replication studies, **Section 2** provides summaries of efforts to improve and understand the replicability of clinical research and **Section 3** outlines the role of replicability in the Food and Drug Administration (FDA) regulatory pathways.

Section 1: Replicability Definitions and Methods

1.1 Key Definitions

Discussions both in academic communities and around the kitchen table often involve many terms which can be interchangeable or represent distinct types of work. Strict definitions serve us less well than clear communication about what, in the given situation, are the planned tasks and the threshold for replicability success or failure. Rather than relying on a single term to communicate a complex set of objectives, the following questions can serve as key points to ask and answer (**Figure 2**).

Figure 2: Key Clarification Questions and Implied Scope of Replication


Collecting new data?
YES
NO
Emulating the previous study sample and protocol?
Testing alternative model specifications and assumptions?
YES
NO
YES
NO
Verifying the effect size
Studying the same outcome or population?
Checking robustness
Checking procedural or coding errors
YES
NO
Verifying effect size with alternative sample size and/or protocol
Generalizing to new populations or outcomes

There are several papers which aim to categorize replication efforts in thoughtful and distinct categories which in addition to the questions posed in **Figure 2** can clarify the type of replication of interest. Goodman and coauthors propose three terms to categorize replication efforts across the dimensions of methods, results, and inference (7).

Methods Replicability

Methods replicability implies that the same procedures could be exactly repeated. The original study provides sufficient data and protocol details for an independent researcher to follow all analytic or experimental steps with the starting point of the original data.

Results Replicability

Results replicability is defined as performing the same study protocol with new, independent data and finding the same (or sufficiently similar results).

Inference Replicability

Inference replicability requires that independent research or reanalysis yields the same conclusions. For instance, it is possible that even with consistent or stable results, there is disagreement about the interpretation or meaning of the results.

The distinction between methods, results, and inference are useful categorizations to consider when designing a replication study. Another categorization system developed by Michael Clemens dichotomizes studies into replication or robustness categories primary from the perspective of economic research (8). Clemens distinguishes replication from robustness on the basis of whether parameters are drawn from the same (replication) or different (robustness) sampling distribution (**Table 1**).

Table 1: Clemens (2017) Proposed Standard of Classifying Any Study as a Replication

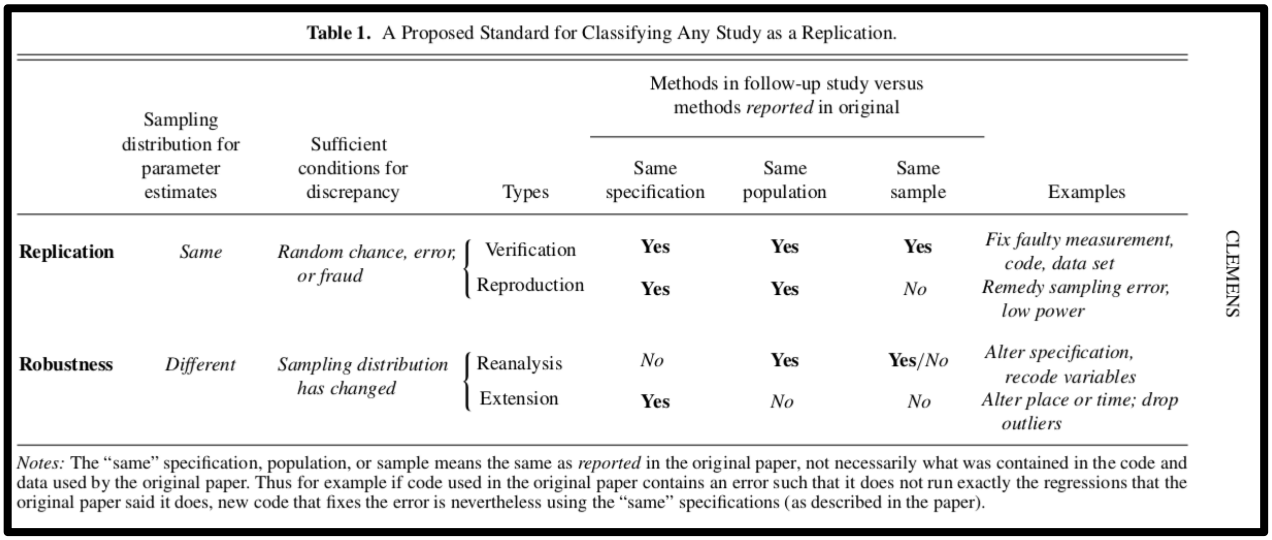


Table 1. A Proposed Standard for Classifying Any Study as a Replication.

				Methods in follow-up study versus methods *reported* in original			
	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	**Yes**	**Yes**	**Yes**	*Fix faulty measurement, code, data set*
			Reproduction	**Yes**	**Yes**	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	**Yes**	**Yes**/*No*	*Alter specification, recode variables*
			Extension	**Yes**	*No*	*No*	*Alter place or time; drop outliers*

Notes: The "same" specification, population, or sample means the same as *reported* in the original paper, not necessarily what was contained in the code and data used by the original paper. Thus for example if code used in the original paper contains an error such that it does not run exactly the regressions that the original paper said it does, new code that fixes the error is nevertheless using the "same" specifications (as described in the paper).

Yet, even within well-defined and structured systems for categorization there is still a need for clear communication around what is considered a replication. This is well articulated by Clemens:

> "If the original sample contains data from one African country but asserts that its results apply to "Africa," then a follow-up study in a different African country could be said to represent the same population by the original article's criteria; but if the original study asserts validity only for the original country, a follow-up in a different country is not sampling the "same" population. The standard proposed here is that it is incumbent upon the author claiming a failed replication to demonstrate that the discrepancy is material, and to offer persuasive reasons why the population should be considered the "same" population."

In addition to having consistent frameworks for thinking and communicating replication study types, it remains useful to know the technical definitions of terms, even if the terms are often used interchangeably in practice (7–9). A comprehensive resource is the National Science Foundation (NSF) subcommittee on replicability in science which provides definitions and NSF internal recommendations (10). The NSF focuses specifically on reproducibility, replicability, and generalizability:

Reproducibility

Is the baseline requirement for an informative study and is defined by the NSF as “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” (10). The burden of this standard varies by the type of research. In some studies, providing raw data files and the processing and analytic code will be sufficient. However in lab-based clinical work, there may be more uncertainty if the original inputs are no longer available (records rather than the underlying clinical samples or animal vectors as one example) (7). The NSF provides one recommendation in this domain:

Recommendation 1: “Each report of research supported by NSF should be accompanied by detailed documentation on procedures to enable an independent researcher to reproduce the results of the original researcher. A report of what these archives contain and how they are accessible should be required in a project’s Final Report and in descriptions of “Prior NSF Funding” in proposals seeking new support.”

Replicability

Replicability is defined by the NSF subcommittee as “the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected” (10). The NSF subcommittee includes two recommendations for improved replicability:

Recommendation 2: “NSF should sponsor research that evaluates various approaches to determining whether a finding replicates and to assess which approach(es) under which circumstances are the most helpful for reaching valid conclusions about replicability.”

Recommendation 3: “To permit assessing replication in various ways, NSF should encourage researchers to report associations between variables using different metrics (e.g., standardized and unstandardized coefficients, effect sizes, odds ratios) and indicating precision of estimates (with standard errors) and to assess the statistical significance of findings using these different methods.”

Generalizability

The NSF defines generalizability as conditions when “the results of a study apply in other contexts or populations that differ from the original one” (10). When study results from one study are not observed in a subsequent study there may be multiple explanations. One might be that the first study results were not successfully replicated, new data was collected but the initial findings were not verified. Another explanation may be that the first study results were not generalized—the effect of the treatment or exposure does not have the same effect in the new data (new population, new context) as it did in the original sample population. Lack of generalizability in contrast to lack of replicability may yield important boundaries and theoretical insights to the given study context (10). The NSF recommendation for generalizability is specific to the difference between replicability and generalizability:

> **Recommendation 4:** “NSF should sponsor research that identifies optimal procedures for practically assessing all types of generalizability of findings (e.g., from a set of study participants to a population, from one set of measures to other measures, from one set of circumstances to other circumstances) and differentiating lack of generalizability from failure to replicate.”

The NSF full report includes further refinement of the terms described above as well as insights and recommendations pertaining to confirmation bias and cumulative evidence. In academic work focused on the classification and definition of replication, terms which are often omitted are exploratory and confirmatory. These terms are useful and fundamental to the transition from pre-clinical to clinical research. When we talk about replicability in clinical research, beyond understanding the scope of replication work which has already been discussed in detail above, it’s important to understand the nature of the current evidence and how future studies will approach building on the evidence base.

Exploratory vs. Confirmatory Studies

In pre-clinical research, scientists develop pathophysiological theories mainly through exploratory studies. This type of study may involve molecular and cellular analyses which are flexible and adaptive (11). Hypothesis may change over time and the use of inferential statistics may or may not apply. The overarching aim of this type of research is understanding diseases and the most promising strategies to mitigate disease. By comparison confirmatory studies build on exploratory studies and/or previous confirmatory work with *a priori* hypotheses aimed at understanding (and advancing) specific therapeutic strategies by demonstrating efficacy in clinically relevant outcomes (11). Kimmelman et al. succinctly summarize the aim of exploratory compared to confirmatory studies as:

> “Exploratory studies should place a premium on sensitivity (i.e., detecting all strategies that might be useful), confirmatory studies should be more concerned with specificity (i.e., excluding all strategies that will prove useless in clinical trials).”

Confirmatory trials build off the evidence available. They can vary from replicating earlier studies to efforts much more expansive – moving from an animal to human model or shifting from a surrogate to true clinical outcome. The FDA gives the following guidance to industry on confirmatory trials (12):

> “A confirmatory trial is an adequately controlled trial in which the hypotheses are stated in advance and evaluated. As a rule, confirmatory trials are necessary to provide firm evidence of efficacy or safety.”
>
> “Confirmatory trials are intended to provide firm evidence in support of claims; hence adherence to protocols and standard operating procedures is particularly important.”
>
> “Firm evidence in support of claims requires that the results of the confirmatory trials demonstrate that the investigational product under test has clinical benefits.”

None of these quotes from the FDA are specific to replication. They instead define pre-specified experimentation. Importantly, all pre-specified hypothesis based studies are (1) synthesizing the information we know now and (2) testing to see if that understanding of the world holds. Confirmatory trials are a form of replication.

1.2 When Have We Successfully Replicated?

There is not consensus on when a replication effort is successful. As soon as an effort deviates from simple coding checks, the waters muddy. For example, in one of the most widely known and published efforts in replicability research, the Open Science Collaboration published replication results for 100 psychology studies published in 2008 in three top psychology journals: Psychological Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology (13). Their findings were dramatic, with only 36% to 47% of studies being successfully replicated. The authors recognized that there was no standard method to determine successful replication, instead they used a combination of methods including subjective assessment, significance and p-values, correlation coefficients between effect estimates, and meta-analysis. When replicability was defined as both the original and replication study having significant p-values, the replication success was 36%. When a meta-analysis of the effects what used (threshold was if the meta-analysis CI included 0), the replication success probability was 47%. **Figure 3** shows the results of the Open Science Collaboration study when replicability was estimated with p-values (**Panel A**) compared to the correlation coefficients of effect estimates (**Panel B**).

Figure 3: Open Science Collaboration – Psychology (2015)

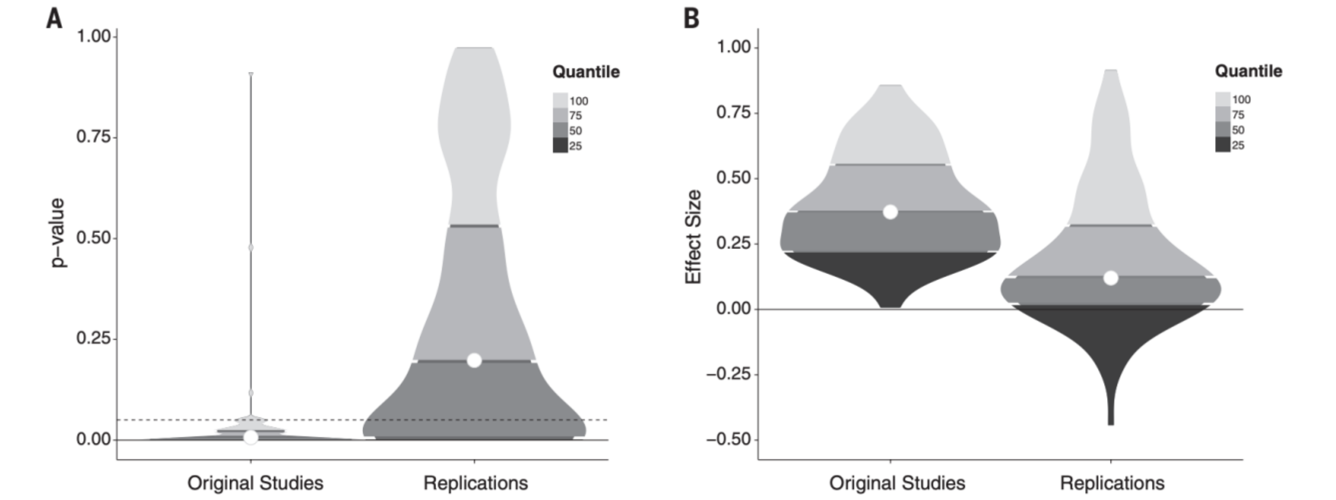


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Despite the authors best efforts to provide a nuanced and complete picture of results replicability, even this highly protocol-driven study received methodological criticism (14). A technical comment by Gilbert et al. (2016) argued that after corrections for error (both sampling and random), power (only one replication study), and bias (original authorship endorsement vs. not) the expected replication success probability is capped at 66%. The analysis from Gilbert et al. implies that the success rate of the Open Source Collaboration was actually reasonably high when adjusting for error, power, and bias. A later paper suggested that 77% of the replication results were successful when benchmarked using the 95% predictive interval (15). The discussions did not end there.

Anderson et al. (2016) published a response to the technical comment which critiqued many of the assumptions made by Gilbert et al. The following quote from the response is emblematic of the state of the field (16):

> “What counts as a replication involves theoretical assessments of the many differences expected to moderate a phenomenon. OSC2015 defined (direct) replication as “the attempt to recreate the conditions believed sufficient for

> obtaining a previously observed finding." When results differ, it offers an opportunity for hypothesis generation and then testing to determine why. When results do not differ, it offers some evidence that the finding is generalizable. OSC2015 provides initial, not definitive, evidence—just like the original studies it replicated."

Methodological development and consensus building around best practices in an ongoing need. Qualitative assessment may have the advantage of focusing in on clinically meaningful differences or congruencies rather than statistical thresholds. However, qualitative assessments are subjective and may be difficult to generalize across settings. Consistent p-values as a criterion is straight forward and intuitive but has serious flaws. The difference between significance and non-significance is not itself necessarily significant as Gelman and Stern so eloquently discuss in their American Statistician piece (17). Confidence and prediction intervals my better account for sampling variation but can't address all of the issues brought up by Gilbert et al. Another option is the predictive value of the original study. For instance, measures like positive predictive value or acceptable false positive rate are common descriptors of diagnostic or screening test results and can be applied to research findings (18). Meta-analytic techniques which summarize the cumulative evidence rather than relying on pairwise comparisons are particularly promising but require a full understanding of the known evidence beyond any one study.

Section 2: Efforts in Clinical Reproducibility & Replicability

2.1 National Institutes of Health

In 2014, the Director of the National Institutes of Health (NIH), Francis Collins along with his Deputy Director, Lawrence Tabak, proposed actions by the NIH to improve the reproducibility of research (19). The first proposal was strengthening training programs in reproducibility and transparency in findings to be piloted on NIH intramural postdoctoral fellows and later disseminated. As of November 2018, there are publically available training modules on the NIH website ranging in topic from the lack of transparency, to blinding and randomization, and sample size. There is also a page for publications related to reproducibility with NIH authors.

The next proposed policy was a checklist to ensure grant applications were evaluated in a standardized format to confirm key design elements like randomization and blinding. At publication, a pilot study was already underway to evaluate a policy of assigning one reviewer to check on grant proposal characteristics of the scientific premise. The article also states that measures will be taken to reduce bias in grant review with a link to the Diversity Working Group Subcommittee on Peer Review.

Finally, Collins and Tabak focus on data access through the use of a Data Discovery Index (DDI). The importance of this system is the ability to associate authorship with research data. A helpful summary of work funded by the NIH towards the development of a Data Discovery Index is available online. The launch of PubMed Commons an online forum to discuss studies available through the PubMed system was also announced. PubMed Commons was later discontinued in March of 2018 due to low participation rates. All existing comments have been archived and can be accessed through the NCBI.

Along with the previously outlined recommendations of the National Science Foundation to improve replicability, the NIH is another key stakeholder to engage. With four years of time between the writing of the piece by Collins and Tabak and the writing of this white paper, proposed policies could be paired with actions. There are some stories of success, like the online training videos, and some policies, like PubMed commons, that will not be a long-term resources.

2.2 The Reproducibility Project: Cancer Biology

The Reproducibility Project is an effort to replicate 50 preclinical studies published between 2010 and 2012 in high profile journals (20). A full list of the articles selected is available online (Database of Articles) as well as a summary of the article selection criterion (Inclusion Criterion). For articles selected, independent labs were contracted to perform the replication experiment after the replication protocol had been peer reviewed and published (20).

The results from this effort have been mixed. Conducting the replications has taken more time and effort than originally anticipated – of the 50 initial studies only 5 have published results (21–25). Those studies, beyond demonstrating the labor intensity of replication, also demonstrate how difficult interpreting the results can be. Of the 5 studies, two reported successful replications of key results, one reported failure to replicate, and two studies were inconclusive. An editorial on the findings urged caution when interpreting any single result, emphasizing the importance of repeated replications and meta-analysis across replication studies:

> "We will publish more Replication Studies over the months ahead and, at the conclusion of the project, a meta-analysis of all the studies (Errington and Nosek, 2017). While we wait for this, it is important not to over interpret the results. Already it is clear that nuanced interpretations are necessary, not black and white conclusions about which studies reproduced and which did not. It is also clear that this approach to testing reproducibility remains an experiment, with advantages and disadvantages, including the fact that it sometimes yields results that cannot be interpreted"

The Reproducibility Project is a collaboration between the Center for Open Science and the Science Exchange. The Center for Open Science is a non-profit organization with a mission of "increasing openness, integrity, and reproducibility in research". They provide an open source project management system in addition to other services which can act as a repository for the entire life cycle of research from idea development through publication. There is documentation capacity and the ability to register work. The Science Exchange is a research outsourcing service which connects scientists with projects. The Science Exchange can be used to contract an independent verification of study results.

The Prostate Cancer Foundation-Movember Foundation have also partnered with the Center for Open Science and the Science Exchange to fund replications of key studies in the field of prostate cancer research. An initial list of 20 studies selected for replication is available on the project webpage.

Section 3: Replicability in FDA Regulatory Pathways

This section focuses specifically on the regulation of drugs. U.S. Code [21 USC 321(g)(1)] defines a drug as "a substance which exerts an action on the structure or function of the body by chemical action or metabolism and is intended for use in the diagnosis, cure, mitigation, treatment or prevention of disease" (26). This white paper focuses specifically on two approval pathways, the standard approval process as well as accelerated approval. This section is not intended as a comprehensive review of all potential approval pathways and processes, rather, the intention is to consider how regulatory requirement confirmatory trials fit in the conversation of replicability in science.

3.1 Standard Review

If preclinical studies are promising, a drug sponsor will submit an Investigational New Drug application (IND). An IND formally engages the FDA and indicates interest in human subject testing to substantiate the safety and efficacy of the drug-indication(s) proposed. There are three phases of human testing: Phase I, II, and III. **Table 3** summarizes the typical characteristics of each study phase. If a drug eventually receives approval and goes to market, there is an additionally Phase IV post-marketing monitoring. For further study, a comprehensive resource to understand the regulation of drugs is the Pisano and Mantus book "FDA Regulatory Affairs" which provides detail into the IND application and clinical research phases (26). The

FDA has also developed a helpful diagram of the approval process which outlines preclinical, clinical, New Drug Application (NDA) review, and post-marketing FDA approval steps.

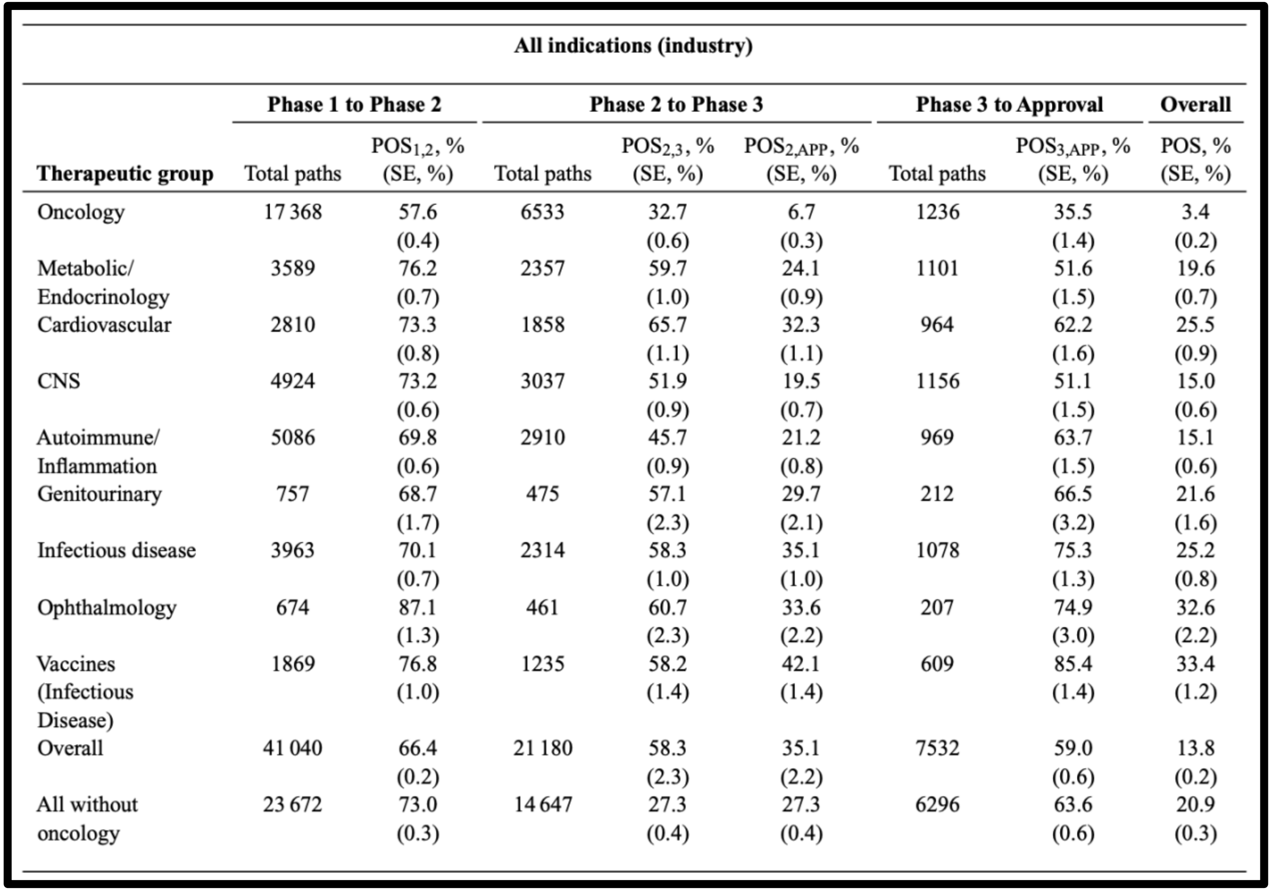
Table 3: Clinical Study Phases

	Study Size	**Key Outcomes**
Phase I	<100; healthy individuals	Toxicology Metabolism Pharmacologic actions
Phase II	100s; w/ treatment indication	Dosing Side effects Safety **Effectiveness**
Phase III	1000s; w/ treatment indication	Safety Drug combinations **Effectiveness**

Effectiveness information from Phase I is minimal, the test subjects are typically healthier than the intended treatment population, the sample sizes are small, and the time frame of observation is limited (often less than one year). However, starting with Phase II, study results should provide information on the safety and efficacy of the drug. With this is mind, drug-indications that make it to Phase III, build on the hypothesis established with data from Phase II, that the drug-indication is more likely than not to be effective and safe, a confirmatory trial.

The success rate from one phase to the next and in particular from Phase III to approval is an empiric context for studying and testing replicability. Further post-market monitoring and the rate of black-box warnings and market withdrawal are important indications of the validity of Phase III results. Wong et al. (2018) have recently estimated the rate of success overall and by phase transition disaggregated by therapeutic group. The overall success rate from Phase I to approval is as low as 3.4% in oncology drugs to as high as 33.4% for vaccines (27). **Table 4** below includes estimates from Wong et al. of successful phase transition by therapeutic category. In the paper, the authors also compare their results to several previous studies estimating the probability of success in drug advancement and final approval (available in Table 1 – Wong et al.).

Table 4: Probability of Success (Wong et al. 2018) by therapeutic group (Jan 2000 – Oct 2015)



All indications (industry)

	Phase 1 to Phase 2		Phase 2 to Phase 3			Phase 3 to Approval		Overall
Therapeutic group	Total paths	$POS_{1,2}$, % (SE, %)	Total paths	$POS_{2,3}$, % (SE, %)	$POS_{2,APP}$, % (SE, %)	Total paths	$POS_{3,APP}$, % (SE, %)	POS, % (SE, %)
Oncology	17 368	57.6 (0.4)	6533	32.7 (0.6)	6.7 (0.3)	1236	35.5 (1.4)	3.4 (0.2)
Metabolic/ Endocrinology	3589	76.2 (0.7)	2357	59.7 (1.0)	24.1 (0.9)	1101	51.6 (1.5)	19.6 (0.7)
Cardiovascular	2810	73.3 (0.8)	1858	65.7 (1.1)	32.3 (1.1)	964	62.2 (1.6)	25.5 (0.9)
CNS	4924	73.2 (0.6)	3037	51.9 (0.9)	19.5 (0.7)	1156	51.1 (1.5)	15.0 (0.6)
Autoimmune/ Inflammation	5086	69.8 (0.6)	2910	45.7 (0.9)	21.2 (0.8)	969	63.7 (1.5)	15.1 (0.6)
Genitourinary	757	68.7 (1.7)	475	57.1 (2.3)	29.7 (2.1)	212	66.5 (3.2)	21.6 (1.6)
Infectious disease	3963	70.1 (0.7)	2314	58.3 (1.0)	35.1 (1.0)	1078	75.3 (1.3)	25.2 (0.8)
Ophthalmology	674	87.1 (1.3)	461	60.7 (2.3)	33.6 (2.2)	207	74.9 (3.0)	32.6 (2.2)
Vaccines (Infectious Disease)	1869	76.8 (1.0)	1235	58.2 (1.4)	42.1 (1.4)	609	85.4 (1.4)	33.4 (1.2)
Overall	41 040	66.4 (0.2)	21 180	58.3 (2.3)	35.1 (2.2)	7532	59.0 (0.6)	13.8 (0.2)
All without oncology	23 672	73.0 (0.3)	14 647	27.3 (0.4)	27.3 (0.4)	6296	63.6 (0.6)	20.9 (0.3)

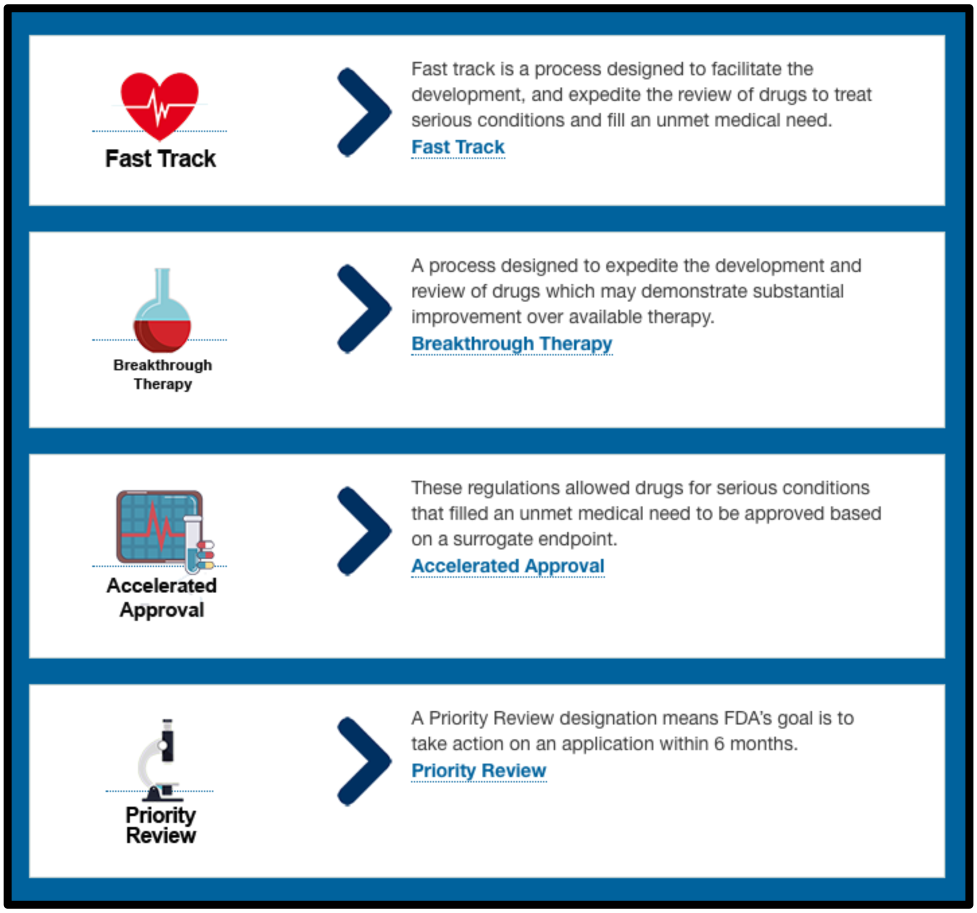
Black-box warnings and drug withdrawals indicate safety concerns which were not apparent in the approval process. Black-box warnings are included in the medical label and can describe severe side effects, dosing information, and potential drug interactions. Drug-indication withdrawals remove a drug from the market for that indication. One study which analyzed New Drug Applications (NDAs) for the period 1975 to 2009 (N=748) found that new drugs have a one-in-three chance of receiving a black box waring or a withdrawal within 25 years of approval, with drugs approved after the 1992 FDA reforms having higher observed rates of black-box warnings and withdrawals (28).

Both the success of a drug-indication from one phase of study to the next, as well as drug-indication withdrawals or black-box warnings are important contexts to research and improve replicability. When drugs fail to progress through the approval pipeline, the later-stage confirmatory trials have failed to replicate earlier findings and hypotheses. When drug withdrawals or black-box warnings are issued, observational or post-market controlled trials failed to replicate the cumulative evidence that lead to the initial drug approval decision.

3.2 Accelerated Approval

In 1992 the FDA introduced the accelerated approval pathway among other reforms (26). The intent of the accelerated approval pathway is to get promising therapies to high-need patient populations faster. Accelerated approval is one of several expedited review options which have detailed descriptions on the FDA's website. **Figure 4** includes the high-level differences between expedited review options.

Figure 4: FDA Expedited Approval Options (source: FDA)



Fast Track

Fast track is a process designed to facilitate the development, and expedite the review of drugs to treat serious conditions and fill an unmet medical need.
Fast Track

Breakthrough Therapy

A process designed to expedite the development and review of drugs which may demonstrate substantial improvement over available therapy.
Breakthrough Therapy

Accelerated Approval

These regulations allowed drugs for serious conditions that filled an unmet medical need to be approved based on a surrogate endpoint.
Accelerated Approval

Priority Review

A Priority Review designation means FDA's goal is to take action on an application within 6 months.
Priority Review

Regulatory decisions in the FDA's accelerated approval pathway are typically made on surrogate or auxiliary endpoints. Of the pre-approval studies informing FDA accelerated approvals from 2009 to 2013, no studies measured overall survival as a primary endpoint (29). Methods to validate surrogate or auxiliary endpoints are well developed but in their most robust form, require randomized controlled trial data (30–33). In the accelerated approval context 84% of auxiliary endpoints had no robust validation studies and for the subset of endpoints with validation studies only 14% (N=4) were considered valid based on international guidelines of R-squared values (34,35).

The FDA can explicitly define the scope of required post-market studies for drug-indications which receive accelerated approval through the FDA Amendments Act of 2007.

However, despite the FDA's authority to set benchmark deadlines and impose fines if post-approval studies are not completed on schedule, delays remain a pervasive issue (29,36). For example, a drug-indication approved in 2009 for multiple sclerosis (Gilenya) has not completed the required post-approval trial to determine efficacy and safety for lower dosages due to recruitment issues while the manufacture has reported over 2.8 billion USD in sales (36).

Of the total 614 post-approval requirements from 2009 and 2010, 20% had not been started by September 2015, at least 5 years post-approval (36). In a study of drug-indications receiving accelerated approval in the time period 2009 to 2013, by April 2017 the post-approval requirements were completed in 10 of 24 indications (42%) (29). Of the total 18 available post-approval studies, only 1 used overall survival at the primary end point (29). The one study that did include overall survival was for the treatment of glioblastoma with Avastin. Beyond the concerns about delays in post-approval confirmatory studies, the quality of evidence produced by these studies is important. Drug-indications are approved and clinically prescribed for years with the assumptions that efficacy will be confirmed, not on surrogate endpoints, like disease response, but on primary outcomes like overall survival and improved quality of life.

In contexts where evidence is used in decision making, verification with confirmatory trials is extremely important. The replication conversation gets to the heart of that point. There are meaningful justifications for the accelerated approval framework. It provides faster access to potentially beneficial drugs to patients with few alternative options. When post-approval evidence is delayed or fails to include clinically meaningful endpoints, potential inefficiencies are introduced. Patients may be receiving care with little to no benefit, health systems may be reimbursing for ineffective drugs, and the regulatory requirements themselves create costly trials without generating information to improve clinical decision making.

The NSF and the NIH have both reflected on replicability and reproducibility within the scope of their organizational work. Both institutions have the ability to use their funding power to set policies which improve the replicability of research. The FDA similarly has the power to set policy regarding replicability through the standards of confirmatory trials. Particularly in the case of accelerated approval where the FDA can stipulate the time frame, endpoints, and impose fines there is incredible opportunity for policy reform.

Conclusion

Replication encompasses many types of efforts from reviewing code to multi-center randomized controlled trials. The success or failure of replication studies can be determined using varied and sometimes inconsistent methods. And despite large efforts to produce data on the replicability of clinical research, we still know very little. These efforts take time, money, and have high potential opportunity costs. However, a rich data source which reveals high-level replicability of clinical research is already available through drug approval phase-specific success rates. When drugs fail to progress from one phase to another, they fail to replicate earlier results. This white paper is intended to illustrate and reflect on each of these points within the context of clinical research.

The term replicability often evokes the idea that someone is attempting to perfectly repeat a protocol. Ultimately, we care about the veracity of results because we act on those results. Many types of studies which can be more or less comprehensive with more or less fidelity to earlier research can help us gain intuition and confidence that we are making the right choices (or not) based on the cumulative information we know. With a broader understanding of the types of studies that can qualify as replication efforts, it becomes increasingly important to strategically pursue a high-value subset of study results. The Cancer Biology Reproducibility Project demonstrates how time intensive and costly replication studies can be.

Given the vast quantity of research produced, replication prioritization based on the importance the effect estimates on decision making is imperative. The link between study results, down-stream decision making and the consequences of those decisions, whether health or economic, is an underdeveloped and discussed area in the field of replicability. By linking pre-clinical and clinical research results with downstream patient outcomes, there is potential to formalize the relationship between early evidence, decision making under uncertainty and the consequences of decision making with uncertainty – the patient outcomes.

Works Cited

1. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011 Sep;10(9):712.

2. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012 Mar 28;483:531–3.

3. Van Norman GA. Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. JACC Basic Transl Sci. 2016 Apr 1;1(3):170–9.

4. Arrowsmith J. Trial watch: Phase II failures: 2008–2010. Nat Rev Drug Discov. 2011 Apr 29;10:328–9.

5. Sasich LD, Sukkari SR. The US FDAs withdrawal of the breast cancer indication for Avastin (bevacizumab). Saudi Pharm J. 2012 Oct 1;20(4):381–5.

6. Baker M. 1,500 scientists lift the lid on reproducibility. Nat News. 2016 May 26;533(7604):452.

7. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016 Jun 1;8(341):341ps12-341ps12.

8. Clemens MA. The Meaning of Failed Replications: A Review and Proposal. J Econ Surv. 2017 Feb 1;31(1):326–42.

9. Hamermesh DS. Viewpoint: Replication in economics. Can J Econ Can Déconomique. 2007 Aug 1;40(3):715–33.

10. Bollen K, Cacioppo J, Kaplan R, Krosnick J, Olds J. Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. Arlington, VA.: National Science Foundation; 2015. (Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences).

11. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. PLoS Biol [Internet]. 2014 May 20 [cited 2018 Nov 12];12(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4028181/

12. FDA: Guidance for Industry E9 Statistical Principles for Clincal Trials [Internet]. U.S. Department of Health and Human Services, Food and Drug Administration; 1998 Sep. Available from: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf

13. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015 Aug 28;349(6251):aac4716–aac4716.

14. Gilbert DT, King G, Pettigrew S, Wilson TD. Comment on "Estimating the reproducibility of psychological science." Science. 2016 Mar 4;351(6277):1037–1037.

15. Patil P, Peng RD, Leek JT. What should we expect when we replicate? A statistical view of replicability in psychological science. Perspect Psychol Sci J Assoc Psychol Sci. 2016 Jul;11(4):539–44.

16. Anderson CJ, Bahník Š, Barnett-Cowan M, Bosco FA, Chandler J, Chartier CR, et al. Response to Comment on "Estimating the reproducibility of psychological science." Science. 2016 Mar 4;351(6277):1037–1037.

17. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. Am Stat. 2006 Nov 1;60(4):328–31.

18. 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 Oct 11;490(7419):187–91.

19. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature. 2014 Jan 30;505(7485):612–3.

20. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. eLife [Internet]. [cited 2018 Nov 11];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270077/

21. Aird F, Kandela I, Mantis C, Reproducibility Project: Cancer Biology. Replication Study: BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Glass CK, editor. eLife. 2017 Jan 19;6:e21253.

22. Horrigan SK, Reproducibility Project: Cancer Biology. Replication Study: The CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors. Massagué J, editor. eLife. 2017 Jan 19;6:e18173.

23. Horrigan SK, Courville P, Sampey D, Zhou F, Cai S, Reproducibility Project: Cancer Biology. Replication Study: Melanoma genome sequencing reveals frequent PREX2 mutations. Davis RJ, editor. eLife. 2017 Jan 19;6:e21634.

24. Kandela I, Aird F. Replication Study: Discovery and preclinical validation of drug indications using compendia of public gene expression data. eLife [Internet]. [cited 2018 Nov 11];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5245962/

25. Mantis C, Kandela I, Aird F, Reproducibility Project: Cancer Biology. Replication Study: Coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. Pan D, editor. eLife. 2017 Jan 19;6:e17584.

26. Pisano DJ, Mantus D. FDA regulatory affairs: a guide for prescription drugs, medical devices, and biologics. 2nd ed. New York: Informa Healthcare USA; 2008. viii+448.

27. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. Biostatistics [Internet]. [cited 2018 Nov 12]; Available from: http://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxx069/4817524

28. Frank C, Himmelstein DU, Woolhandler S, Bor DH, Wolfe SM, Heymann O, et al. Era Of Faster FDA Drug Approval Has Also Seen Increased Black-Box Warnings And Market Withdrawals. Health Aff (Millwood). 2014 Aug 1;33(8):1453–9.

29. Naci H, Smalley KR, Kesselheim AS. Characteristics of Preapproval and Postapproval Studies for Drugs Granted Accelerated Approval by the US Food and Drug Administration. JAMA. 2017 Aug 15;318(7):626–36.

30. Weir CJ, Walley RJ. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Stat Med. 2006 Jan 30;25(2):183–203.

31. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostat Oxf Engl. 2000 Mar;1(1):49–67.

32. Elston J, Taylor RS. Use of surrogate outcomes in cost-effectiveness models: a review of United Kingdom health technology assessment reports. Int J Technol Assess Health Care. 2009 Jan;25(1):6–13.

33. Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015 Aug;175(8):1389–98.

34. Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Med. 2017 21;15(1):134.

35. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. JAMA Intern Med. 2015 Dec;175(12):1992–4.

36. Woloshin S, Schwartz LM, White B, Moore TJ. The Fate of FDA Postapproval Studies. N Engl J Med. 2017 Sep 21;377(12):1114–7.