Statistical Perspectives on Reproducibility: Definitions and Challenges

Andrea Simkus[1], Tahani Coolen-Maturi[1*], Frank P. A. Coolen[1] and Claus Bendtsen[2]

[1]Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, United Kingdom.
[2]Data Sciences & Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK.

*Corresponding author. E-mail: tahani.maturi@durham.ac.uk;
Contributing authors: andrea.simkus@durham.ac.uk;
frank.coolen@durham.ac.uk; Claus.Bendtsen@astrazeneca.com;

Abstract

Reproducibility is a widely discussed topic, yet many experimental results cannot be confirmed due to factors such as publication bias, poor documentation, and inappropriate statistical methods. A lack of standard definitions for reproducibility and related terms further complicates the matter. This paper reviews the literature on reproducibility, clarifies key terminology by defining five types of reproducibility, and addresses variations in published definitions by considering changes in datasets, labs, and experimental conditions. We explore the causes of low reproducibility in scientific studies and discuss statistical perspectives on quantifying and improving reproducibility. In particular, we propose framing statistical reproducibility as a predictive problem, providing a framework to evaluate and address reproducibility challenges.

Keywords: nonparametric predictive inference, preclinical research, reproducibility, replicability, statistical reproducibility

1 Introduction

Reproducibility [9] is a complex issue, gaining importance and attention in scientific research. *Nature* published a special edition *Challenges in irreproducible research*, dedicated to the problem of researchers not being able to verify results presented in published papers of other scientists [106]. In the literature on the topic of reproducibility there has been a lot of confusion about what the term reproducibility means [64], which will be addressed in this paper.

A better understanding of reproducibility of tests is crucial for pharmaceutical research and development, as a lack of reproducibility contributes to failure rates in drug discovery and development processes, increasing costs, and decreasing efficiency. Begley and Ellis [17] highlighted a systematic problem in preclinical cancer research: the majority of publications in this research area cannot be validated. Scientists at the biotechnology firm Amgen tried to confirm findings of 53 published papers in haematology and oncology by performing replicate experiments. These did not reproduce conclusions in 47 out of 53 studies, even with the attempts to contact the original authors of the articles and to discuss the details of the experiments with them [17]. Errington et al. [52, 53] attempted to carry out replicate experiments based on high-impact papers published in 2010-2012 in the field of preclinical research in cancer biology. A replicate study is a new study, trying to closely imitate the original study. Out of the chosen 193 experiments from 53 papers, they managed to conduct a replicate study for only 50 experiments from 23 original papers. 40% of replications of positive effects and 80% of replications of null effects were successful, according to three or more of five methods of replication assessment, defined by Errington et al. [53].

It is evident that scientists show considerable interest in the topic of reproducibility (or a lack thereof). The number of publications considering reproducibility is large. There is a rich body of literature on reproducibility in pharmaceutical research, particularly in preclinical research [25, 86, 91, 92, 128, 148], which will be discussed in Section 6. In psychology [94, 96, 112, 114, 143, 154], the focus is on the discussion of replicating the outcomes of the original study and the concern about low reproducibility rate (or rather *replicability* rate, as commonly used in psychology). Computer sciences, machine learning and artificial intelligence [32, 66, 119] mainly focus on transparency and sharing of data, code and clear documentation of the whole study. Ioannidis [81] argued that sharing protocols, materials, software, and data provides a sound basis for reproducible data practices. This aspect is also important in chemistry [20, 62], nevertheless, there is also practical advice on how to maximise reproducibility through good laboratory practice and minimising human error.

The purpose of this paper is to provide a review of the literature on reproducibility and highlight important debates on the topic. It is intended for a broad audience, including not only statisticians, but also anyone interested in the subject. By shedding light on the issue of reproducibility, this paper aims to contribute to the ongoing discussion in this field. The authors would like to

note that, although reproducibility is part of the discussion on doing quality research, it does not equate to it. Thus, this paper does not aim to address all aspects of quality research.

Given that there are no standardised definitions for reproducibility and related terms (such as replicability), and that some definitions of reproducibility from the existing literature lack clarity themselves, this review begins by discussing various possible interpretations and definitions of the concept of reproducibility in Section 2. Terms that are related to, or used interchangeably with, reproducibility are also discussed. With the aim to describe the subtleties encountered in the literature, the available definitions are classified into five categories, which we refer to as Type A to Type E. Section 3 outlines the reasons for low reproducibility and reviews suggestions for improving reproducibility as presented in the literature. Section 4 focuses on statistical reproducibility, providing a classification of its definitions and summarising key issues that have been raised. We also briefly address one of the ongoing debates in the reproducibility crisis: whether p-values should be used.

This paper primarily focuses on statistical reproducibility. While much of the literature concerns the validation of test conclusions derived from both the original and replicate experiments, as addressed in Section 5.1, it is also important from a statistical perspective to examine reproducibility in cases where only the original experiment has been conducted. Section 5.2 presents methods for assessing reproducibility in scenarios where only the original experiment is available, suggesting that one approach is to frame reproducibility as a prediction problem, quantify it through the statistical framework of nonparametric predictive inference (NPI). In addition, Section 6 presents a case study on reproducibility issues in preclinical research. The paper concludes with Section 7, which summarises the key points discussed and highlights directions for future research.

2 Definitions of Reproducibility

There is no universally agreed definition for the concept of reproducibility and there are many related terms to reproducibility, such as repeatability, replicability, generalisability, robustness, reliability, open science, transparency, truth [107] and precision [79]. These related concepts are often also not clearly or appropriately defined, some of them are used interchangeably and they are all important for the reproducibility debate. This section presents a summary of definitions for reproducibility used in the existing literature.

Recent overviews of definitions of reproducibility and related terms have been presented by Goodman et al. [64], Barba [14] and Gundersen [66]. Goodman et al. [64] identified that the term *research reproducibility* is not settled both linguistically and conceptually. Barba [14] raised the problem of different groups of researchers using different terminology for the same definition

of reproducibility and related terms. She also mentioned that the terms *reproducibility* and *replication* are often used interchangeably by researchers, which creates confusion and leads to conceptual ambiguity in the literature [14].

This paper will classify the definitions of reproducibility encountered in literature into five categories, Type A to Type E, rather than adhering to precise definitions. The nuances are captured in the 'Reproducibility types tree' in Figure 1. This figure outlines possible considerations that are important for defining reproducibility and related terminology. In describing different types of reproducibility, three key terms are used: data, method and conclusion. Data are "information, especially facts or numbers, collected to be examined and considered and used to help decision-making" [44]. The term method refers to the way the experiment is run. Method encompasses experimental design, data collection method, statistical analysis, software used to analyse the data and programming code. The range of features the method contains differs across different research areas. Conclusion is " a reasoned deduction or inference' [49], conclusion is reached after applying statistical analysis to the data. Next, the five Reproducibility types will be introduced.

Reproducibility Type A: Reproducibility is the ability to follow the analysis of an experiment based on the same data and a clear description of the data and the method.

Stronger version of **Reproducibility Type A:** Experimental conclusions are reproduced if another researcher applied the same analysis to the same data and reached the same conclusions, using the description of the data and the method provided by the original researcher.

Reproducibility Type B: Experimental conclusions are reproducible if same data but a different method of statistical analysis lead to the same conclusion.

Reproducibility Type C: Experimental conclusions are reproducible if new data from a new study carried out by the same team of scientists in the same laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

Reproducibility Type D: Experimental conclusions are reproducible if new data from a new study carried out by a different team of scientists in a different laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

Reproducibility Type E: Experimental conclusions are reproducible if new data from a new study, using a different method of experiment design or analysis, lead to the same conclusion.

All the types of reproducibility rely on a common underlying principle: namely, that the same conclusions 'would be' or have been reached in a reproducible experiment. Reproducibility Type A and Type B do not require new data, and Reproducibility Type C to Type E assume either the existence or

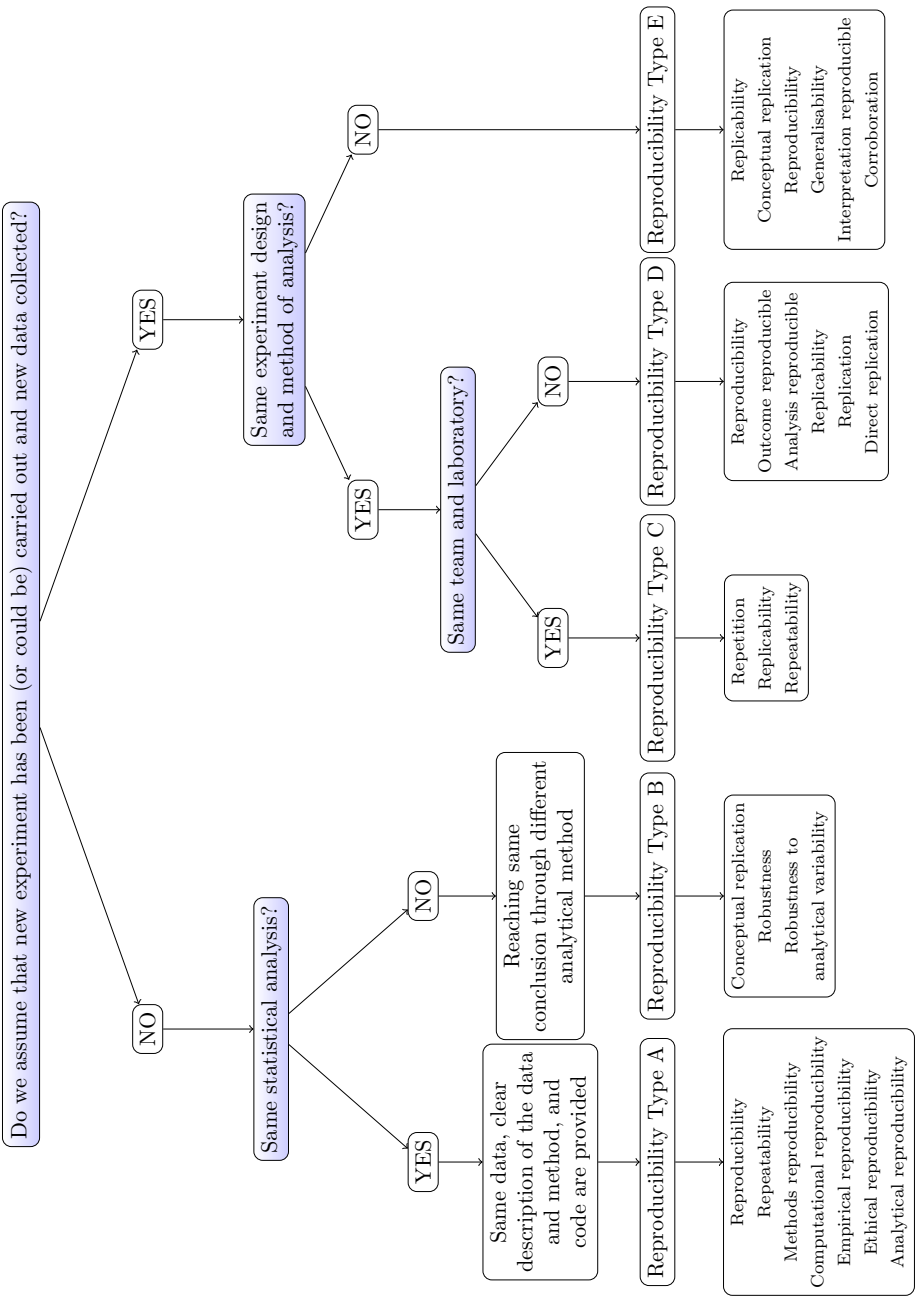

Fig. 1 Reproducibility type tree.

the possibility of existence of new data. The term 'new study' is used in the Reproducibility Types instead of 'replicate study' because it is more general and it does not imply that the follow-up study exactly mimics the original study. This linguistic choice was made, as the term replicate study does not fit well Reproducibility Type E. Throughout this paper, the terms new study and replicate study are used interchangeably.

There is a different debate, distinct from the Reproducibility Types classification, about whether there is a necessity to 'reproduce' (validate) the results by doing the experiment again. This debate deals with the question of whether the same conclusion 'would be' reached if the experiment was carried out again or whether the same conclusion has been reached after the new experiment has been carried out. Reproducibility Type C to Type E do not distinguish between these two options. Reproducibility has been assessed under both scenarios: First, when the new replicate experiment has been performed, which is addressed in Section 5.1. Secondly, when only the original experiment has been performed, and a probabilistic assessment is made about the reproducibility based on the current data and analysis. Section 5.2 will focus on the latter scenario.

In the literature [8, 79, 87, 89, 94, 103], reproducibility has also been defined in terms of *precision* - the closeness of agreement between multiple (two or more) test results obtained under specific conditions, such as same method and same or different test operator. Blackman [23, 24] and Pryseley et al. [8] considered the quantification of this closeness of agreement between test results. In the above mentioned references, related terms employed instead of *precision* are *reproducibility*, *repeatability*, *measurement precision*, *measurement repeatability* and *measurement reproducibility*. However, in these references there is no distinction between the original and the new test. The focus of these studies is on the variability in repeated measurements when one or more elements of the study, such as time, the observer, environment or instruments, are different [103]. We do not believe that *precision* should be considered equivalent to reproducibility. The concept of the original study is crucial to any discussion about reproducibility. Therefore, the assessment of the closeness of agreement is not within the scope of this overview.

Sections 2.1, 2.2 and 2.3 will elaborate on the variety of definitions for reproducibility and related terms available in the existing literature, and classify these terms into five different types of reproducibility. This classification into types aims to clarify different terminologies. It also aims to show inconsistency in terminology used across different publications and unclarity of some definitions. Each reproducibility type will be discussed separately, with the exception of Reproducibility Type C, Type D and Type E, which are discussed in the same section. The reason for this is that there are some definitions that either refer to multiple reproducibility types or it is unclear which of these three types they refer to.

2.1 Reproducibility Type A

In alignment to Reproducibility Type A, the requirement for reproducible research is that the documentation, data and code used for the analysis are available to others, so that they can verify the published results or carry out alternative analyses [118]. Goodman et al. [64] called *methods reproducibility* the ability, rather than necessity, to reach the same conclusion by using the same data and method. Here ability refers to the availability of the data and a clear description of the data and the method, which would allow the researcher to re-enact the analysis. National Science Foundation's [110] definition of *reproducibility* can also be classified as Reproducibility Type A. Similarly, a workshop organised by the National Academies of Sciences, Engineering, and Medicine (NASEM) [107] used the term *repeatability* (also called *empirical reproducibility*), which can be classified as Reproducibility Type A. Donoho [50] used the term *computational reproducibility* without explicitly defining it, nevertheless its use is in alignment to Reproducibility Type A. Peng [119] argued that there is a spectrum of reproducibility. On the lower end of the spectrum is limited code sharing, in the middle section of the spectrum is sharing code and data, and on the upper end of the spectrum is sharing a single file containing both data and code that can execute the full analysis of the data. According to Peng [119], this upper end of the spectrum means full replication, which is the gold standard for reproducibility, as it allows the researcher to carry out the full analyses again [119]. Peng's spectrum of reproducibility does not encompass the term method, however, it is possible that this is because Peng discussed reproducibility in computational science, where code represents the method. Peng et al. [120] defined criteria for reproducible epidemiologic research as the availability of data, method, documentation and accessibility to the software, data, and documentation, which classifies as Reproducibility Type A. Gentleman and Lang [61] define *reproducible research* as research articles which are accompanied with software tools allowing readers to reproduce the paper results and further use the computational methods presented in the paper. Their definition can also classify as Reproducibility Type A.

Stodden [144] divided Reproducibility Type A into *empirical reproducibility*, which requires appropriate reporting standards and documentation of the physical experiment, and *computational reproducibility*, which requires accommodating the use of computation technology in the reporting and scientific practice. *Ethical reproducibility* [5], for which it is imperative to transparently report ethical challenges and methods of resolution of them in studies in biomedical research, also falls into the category of Reproducibility Type A. Thus there is a reasonable body of work that adopts definitions, which can be classified as Reproducibility Type A.

Reproducibility Type A leads to better transparency in research. We agree that careful documentation of an experiment should be part of creating reproducible research. We expect all research to have data, method, and code available upon request, but given the amount of the literature on definitions which classify as Reproducibility Type A, this is likely not the case. One reason

is the lack of incentives for researchers, particularly when such documentation is not explicitly required by journals or funding bodies. In computer sciences, Collberg et al. [32] conducted a study to determine whether 613 papers (from eight Association for Computing Machinery conferences and five computer science journals) presented reproducible research. Only papers, for which Collberg et al. [32] were able to obtain code and execute it, were labeled as reproducible research - reproducible in accordance with Reproducibility Type A. These were 102 out of 613. Collberg et al. [32] did not verify the accuracy of the published results. They provided an elaborate list of reasons why researchers did not provide code after email correspondence, examples of these reasons were: bad backup practices, the student who programmed the code left the research institution, and the code being an intellectual or commercial property.

A stronger version of Reproducibility Type A is presented by Benjamini in the proceedings of NASEM [107]. He defined *reproducibility* of the study as reaching the same conclusions after performing the same analysis on the study's raw data. *Reproducibility* in Errington et al. [52], Stevens [143] and Nosek et al. [109], and a consensus study report by the National Academies of Sciences [108] can also be classified as the stronger version of Reproducibility Type A. Botvinik-Nezer and Wager [28] called the stronger version of Reproducibility Type A *analytical reproducibility*. An article in *Biostatistics* is defined as reproducible if the Associate Editor for Reproducibility executed the code on the provided data and reproduced the results given in the article [118], which is also an example of the stronger version of Reproducibility Type A.

2.2 Reproducibility Type B

The core feature of Reproducibility Type B is that experimental conclusions are reproducible if the same data but a different method of data analysis were used to reach the same conclusion. While Reproducibility Type B is not a widely discussed kind of reproducibility, a reference to it can be found in Stahel [142], Goodman et al. [64], Errington et al. [52] and Botvinik-Nezer and Wager [28]. Stahel [142] discussed *conceptual replication*: where different analytical methods are used on the same data to re-examine conclusions of a study, which can be categorised as Reproducibility Type B. Errington et al. [52] used the term *robustness* for using alternative strategies on the same data, which also classifies as Reproducibility Type B. Similarly, Botvinik-Nezer and Wager's [28] terminology *robustness to analytical variability* fits with Reproducibility Type B. According to Possolo [123], different statistical methods - models and data analysis methods, including data reduction - can lead to different conclusions, when the same data is analysed.

Goodman et al. [64] presented *inferential reproducibility* which leads to similar conclusions from "an independent replication of a study or a re-analysis of the original study." The latter part of their definition could either refer to Reproducibility Type B or stronger version of Reproducibility Type A, depending on whether the re-analysis uses the same method as the original one

did. The former part requires new data and new analysis, which is in alignment with Reproducibility Types C, D and E, which are discussed in Section 2.3.

The literature does not clearly specify what is meant by a 'different method of statistical analysis' in Reproducibility Type B, nor is it clear how different the method should be. More consideration is needed to clarify this distinction. Ensuring that the statistical analysis is appropriate and suitable is an important aspect to address. It is also important to highlight that the negative version of Reproducibility Type B, i.e. experimental conclusions being irreproducible due to different statistical reproducibility not leading to the same conclusion as the original statistical reproducibility, has not been considered in the literature. This negative version of Reproducibility Type B would be absurd, given that in many cases, there is some statistical analysis that can lead to a different conclusion than the original statistical analysis. The reason behind the lack of exploration of these two mentioned aspects of Reproducibility Type B may be that reproducibility has been widely discussed by non-mathematicians, and the discussion lacks mathematical rigour.

2.3 Reproducibility Type C, Type D and Type E

A combination of Reproducibility Type C, Type D, and Type E often fits a particular definition. For example, a consensus study report by the National Academies of Sciences, Engineering, and Medicine (NASEM) [108] defined *replicability* as "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." This rather broad definition of replicability includes Reproducibility Types C, D and E. The same definition of *replicability* was adopted by Errington et al. [52], Patil et al. [115] and Stevens [143]. Similarly, Nosek et al. [109] referred to *replication* as using different data to test the reliability of prior findings. The special issue on reproducibility and replicability [31] published in *Statistical Science* also follows NASEM's [108] definition of replicability.

In fact, it can be hard to distinguish under which type or types of reproducibility a particular definition can be categorised. Unclarity of definitions is a problem in the reproducibility debate. An example of an ambiguous and vague definition of reproducibility is Goodman's definition of reproducing the results of investigators in the proceedings of NASEM [107]. This is defined as "finding the same evidence or data, with the same strength." It is unclear how to assess whether the requirement of this definition has been met.

Another example of unclear definitions has already been discussed in Section 2.2: *inferential reproducibility*. It is unclear what Goodman et al. [64] meant by the former part of the definition for *inferential reproducibility*: achieving similar conclusions from "an independent replication of a study." It is unclear whether or not the circumstances of the original and the replicate study were identical or may have varied. If so, then it would classify as Reproducibility Type E. However, this is just a possible interpretation of the definition of *inferential reproducibility*. The definition could also fit with Reproducibility Type C or D. Moreover, this vague definition of reproducibility allows for the

possibility of replicate study leading to considerably different data. In a different literature source, Goodman [107] defined *inferential reproducibility* as "reaching the same conclusions or inferences based on the results," however, this new definition does not yield more clarity.

Furthermore, Goodman et al. [64] defined *results reproducibility* as "obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible." It is unclear whether the definition categorises as both Reproducibility Types C and D or just the latter. The reason for this unclarity is that the term *independent study* has no clear definition and it is often used by researchers, including Goodman et al. [64], without being defined. It is unclear whether in Goodman et al.'s [64] definition of reproducibility, the same team of scientists or a different one has to carry out the experiment. On the other hand, Voelkl et al. [147] provided a clearer definition of reproducibility. They defined *reproducibility* as "the ability to produce similar results by an independent replicate experiment using the same methodology in the same or a different laboratory," which encompasses both Reproducibility Types C and D. A similar definition was used by Richter [127].

The National Science Foundation [110] distinguished three terms: *reproducibility*, *replicability* and *generalisability*, and they saw these as a foundation for robust scientific findings. *Reproducibility* can be categorised as the Reproducibility Type A definition, as stated in Section 2.1; *replicability* is the ability to validate the results of a prior study by collecting new data via the same procedure [110], which fits both Reproducibility Types C and D. *Generalisability* is attained when "the results of a study apply in other contexts or populations that differ from the original one" [110]. Their definition of generalisability is most relevant to Reproducibility Type E.

Jarvis and Williams [86] defined *replication* as obtaining an identical result in an experiment conducted under identical conditions, which is compatible with Reproducibility Type C and possibly also with Reproducibility Type D, as it is unspecified whether the same team of scientists or the same laboratory is necessary. Jarvis and Williams [86] defined *reproducibility* as obtaining a similar result in an experiment conducted "under similar yet different conditions, the latter having the necessary degrees of latitude that reflect a real-world situation," which is most compatible with Reproducibility Type E. It is unclear what the terms 'similar results' and 'similar yet different conditions' exactly mean.

Barba [14] presented another division of terminology: *repeatability*, requiring the same team and the same experimental design, which can be categorised as Reproducibility Type C; *replicability*, requiring a different team and the same experimental setup and fitting in Reproducibility Type D; and *reproducibility*, requiring a different team and a different experimental design, which can be classified as Reproducibility Type E. It is consistent with definitions of *repetition* and *replication* by Atmanspacher and Maasen [9]: *repetition* refers to doing the same experiment by the same team whereas *replication* refers

to situations where different teams carry out the same experiment. However, according to Atmanspacher and Maasen [9], *reproducibility* covers both terms. Gundersen's [66] use of the term *repeatability* also fits with Reproducibility Type C.

Zwaan et al. [154] defined *direct replication* as "studies intended to evaluate the ability of a particular method to produce the same results upon repetition." In this replication, critical elements of the study, such as procedures, samples and measures, are recreated [154]. But only "those elements that are believed necessary for producing the original effect" must be present in the replicate study. This definition is closest to Reproducibility Type D. Furthermore, Zwaan et al. [154] defined *conceptual replication*, which assesses whether an effect extends to a different population. *Conceptual replication* falls under Reproducibility Type E.

Gundersen [66] viewed *reproducibility* in the light of the scientific method. Gundersen's definition of reproducibility requires that a new experiment, mimicking the original experiment by following the documentations from the original researcher, is carried out by another team of investigators [66]. Gundersen's categorisation of reproducibility differs from ours classification of reproducibility definitions. Gundersen [66] categorised reproducibility into four types of reproducibility, which define what type of documentation on the original study was available to the investigators carrying out the new study. Gundersen's four types are: description, code, data, and experiment. The last type encompasses all of the previously named three types. In our view, all details of the original study should be shared in order to replicate the study.

Furthermore, Gundersen [66] categorised reproducibility into three degrees of reproducibility, which he called *outcome reproducible*, *analysis reproducible* and *interpretation reproducible*. The degrees of reproducibility are based on what factors are similar in the original and the replicate experiment. *Outcome reproducible* means that the outcomes in the original and the replicate experiment are the same, thus applying the same analysis leads to the same conclusion. It is vague what outcome means in this context, nevertheless, Gundersen [66] stated that outcomes of some experiments are data. *Analysis reproducible* means that when the same analysis is applied to the new data in the replicate study, the same conclusion is reached. Arguably, both *outcome reproducible* and *analysis reproducible* fall under Reproducibility Type D. Interpretation of the analysis denotes the conclusion made about the study. *Interpretation reproducible* means that neither the outcome nor the analysis have to be the same in the replicate experiment, but the interpretation (conclusion) drawn from the original and the replicate study are the same. Thus, *interpretation reproducible* allows for different statistical analysis and it can be categorised as Reproducibility Type E. Parmigiani [113] discusses replicability of prediction rules which he defines as "obtaining consistent results across studies suitable to address the same scientific prediction question, each of which has obtained its own data." His definition is specific as it refers to prediction rules, which can be applied, for example, in predicting survival of

ovarian cancer patients or screening for tuberculosis [113]. Nevertheless, it is in line with Reproducibility Type E.

Gundersen [66] argued that using different methodology no longer falls under reproducibility, but it is called *corroboration*, as in such cases hypotheses are supported by new evidence. *Corroboration* fits the best within Reproducibility Type E. Gundersen [66] also stated that corroboration refers to theories and hypotheses rather than to experiments. This underlines that there is a considerable disagreement on what constitutes reproducibility.

According to Stahel [142], there are two aspects of a successful replication. One of the two aspects is the *confirmation of conclusions*, which means that a replication study leads to the same conclusion as the original one. This is compatible with Reproducibility Types C and D. Stahel [142] also noted that the concept of reproducibility could be extended to exploring different circumstances, and if the new study leads to the same conclusion as the original study, then this is called *generalisability*, which could be interpreted as using a different method from the original study and can be classified as Reproducibility Type E. Another aspect of successful replication, according to Stahel [142], is *statistical compatibility*, which addresses the question "Is the data obtained in the replication compatible with the data from the original study in the light of the model used to draw inference?" This approach to reproducibility is not discussed elsewhere in the existing literature, and this paper does not classify this as any of the Reproducibility Types because it is not clear what 'data from a new study are compatible with the data from the original study' exactly mean.

Patil et al. [116] recommend that researchers use the R package `scifigure` to visualise their definitions of reproducibility and replicability, facilitating meaningful communication. The visualisations include the following variables: population, question, hypothesis, experimental design, experimenter, data, analysis plan, analyst, code, estimate, and claim. For each variable, the options available are: observed, missing, different value, and incorrectly reported. These visualisations could be particularly useful for discussing Reproducibility Types C, D, and E.

2.4 Summary and Discussion

There is no universally accepted definition of the term *reproducibility*, nor is there clarity around related terms such as *repeatability*, *replicability*, and *generalisability*. Different definitions for the same terms are sometimes used in the literature, and researchers often refer to reproducibility without providing a clear definition. Even when definitions are offered, they are not always precise, and some terms within those definitions remain undefined. For meaningful discussions on reproducibility, it is crucial to clarify the terminology. In this section, we identified five types of reproducibility, along with definitions from the literature that align with these categories. Figure 1 outlined key considerations in defining reproducibility and how they relate to these five types. There

is some overlap among the types, but it is not the aim of this paper to determine which is the "correct" one. Rather, all of the considerations presented are relevant to the reproducibility of scientific findings.

Having reviewed the various definitions and types of reproducibility, it is equally important to consider the goals associated with it. These goals shape how reproducibility is applied and assessed in scientific research. One clear goal is the confirmation of conclusions, though researchers need not limit themselves to this objective. According to Bayarri and Mayoral [16], other goals of reproducibility include the reduction of random error, bias detection, and extension of conclusions. The latter relates to Reproducibility Type E. Goodman [107] expanded this list by adding two additional goals: learning about the *robustness* ("resistance to minor or moderate changes in experimental or analytic procedures and assumptions") and the *generalisability* of results ("true findings outside the experimental frame or in a not-yet-tested situation"). However, *robustness* and *generalisability* may not be entirely new goals but rather clarifications of the broader goal of extending conclusions. Zwaan et al. [154] introduced another role for reproducibility, termed *replication*, to provide more accurate estimates of effect sizes. This goal differs from the others, raising questions about whether reproducibility should focus on estimating effect sizes or whether effect sizes should instead be part of general statistical analysis.

3 Low Reproducibility: Causes and Possible Improvements

There is no universally accepted notion of what low or poor reproducibility means, which is likely linked to the lack of a universal definition for the term *reproducibility*. There are two main approaches to defining low or poor reproducibility: First, it can refer to a poorly described and documented experiment, which prevents another researcher from reproducing the original experiment. This could be done either by using the same data, in alignment to Reproducibility Type A or Type B, or by redoing the experiment and acquiring new data, in alignment to Reproducibility Types C to E. Secondly, low reproducibility can refer to a well described and documented experiment, where a new experiment does not lead to the same findings that were reached in the original experiment. Poor reproducibility can also refer to a combination of these two approaches.

The solutions for improving reproducibility often require adhering to good scientific practice and using appropriate statistical, experimental and documentation methods. The majority of these solutions are not limited to a particular type of reproducibility. Finding solutions to the reproducibility crisis calls for many stakeholders: researchers, public and private institutions, funding bodies, and journals. All these stakeholders can play a vital role in improving reproducibility [107].

3.1 From a statistical perspective

3.1.1 Poor statistical choices

The discussion of statistical reasons for poor reproducibility begins by highlighting the problem of researchers making poor statistical choices. Wrong or unsuitable statistical analysis [12, 19, 57, 108, 124, 144] and poor experimental design [12, 108] are commonly named. This includes the incorrect use of p-values [19], overrelying on p-values [70], inadequate sample sizes [57, 124], low power [144], using inappropriate sampling techniques [144], insufficient knowledge of data-generation mechanisms caused by the use of big data [144], experimental biases [19], statistical biases such as confounding [19], and programming errors [19]. The discussion of the reasons for low reproducibility in the quoted papers is mostly theoretical, and it does not include real-world examples.

More specific reasons for low reproducibility, which only apply to certain experiments, are: examination of weak and complex interactions for data with low signal-to-noise ratio [118], and miscalculation of effect sizes in meta-analyses [1]. Moreover, greater availability of data and more complicated analytical methods lead to a greater risk of false or misleading findings [118], as this increases the risk of an error.

Statistical solutions to problems offered in the literature are: using suitable statistical methods [12, 19, 108], which may include reporting confidence intervals rather than just p-values [70]; using robust designs [12]; acknowledging uncertainties [108] and taking into consideration the sensitivity of estimates for both deviations in the underlying data and model choice [144]. To ensure the appropriate use of statistics, it is important to involve statisticians at all the stages of the experiments or to provide good statistical training to the researchers carrying out the experiments [96, 126]. It is important to teach researchers that statistics is a tool to assess the strength of evidence, rather than to reveal the truth. Even, with this priority, occasional human error is still inevitable.

Berger [19] suggested using Bayesian analysis, and he argued that this statistical framework provides a systematic way of dealing with multiple statistical analyses. Johnson et al. [88] recommend using Bayes factors and posterior model probabilities instead of, or alongside, p-values. Researchers not limiting themselves to either frequentist or Bayesian statistics is desirable. However, this again requires proper statistical training. Stahel [142] encouraged cross-validation, where the dataset is split multiple times into smaller training sets. Model parameters are estimated for each split, and the average performance across all splits is then calculated. Using appropriate statistical analysis has the potential to reduce the incidence of wrong conclusions, which are often caused by technical errors.

3.1.2 Undesired correlations

Stahel [142] named the within laboratory or within group correlation, which is about measurements from the same laboratory being more similar. Apart from the incorrect use of statistical methods, unwanted or unknown correlations could negatively affect reproducibility. Stahel [142] suggested that there may be a correlation between results obtained with short time lags. Carrying out the same experiment in a different laboratory and by a different team of scientists, in line with Reproducibility Type D, can ensure that the conclusions of the experiment are not linked to some of the unknown correlations, such as the within laboratory correlation, which can occur with Reproducibility Type C. One way to reduce the undesired consequences of unwanted correlations is accounting for reproducibility in the design of the experiment. This solution has already been addressed in preclinical research, and this topic will be further discussed in Section 6. Reproducibility Type E avoids some of the pitfalls of unwanted correlations, namely, it tests the findings of the experiment under changed circumstances, which makes the conclusions more robust with respect to the varying conditions. According to Ehm [51], meta-analysis is needed because of the issues of heterogeneity and selection bias [51]. Meta-analysis is a statistical method that combines results of several independent studies [68]. It should not replace replication studies, but it is useful as it can stop researchers from prematurely accepting conclusions. However, Bogomolov and Heller [26] warn that meta-analysis can conclude that there is discovery in cases where there is a statistically significant effect in only one of the studies. Thus, researchers should be carrying out meta-analysis with care.

3.1.3 Within-study selection bias

Related to the undesired correlation is the within-study selection bias. Hutton and Williamson [77] showed, via a meta-analysis on a treatment for incontinence and anthelminth therapy, that selective reporting of outcomes has an effect on the conclusions and recommendations made about treatment. This within-study selection bias is often based on the significance level and the estimates of effect size. However, this selective reporting becomes problematic when meta-analysis is carried out or someone else tries to replicate the experiment. To avoid these problems, all outcomes should be reported, even those that were statistically insignificant.

3.1.4 Missing data

In research, missing data and the lack of documentation of missing data [139] can lead to poor reproducibility. Thus, it is important to report information about missing data, which includes the degree of and statistical assumptions related to missing data, and the practical information on reasons behind missing data. Moreover, it is vital to perform sensitivity analysis to assess robustness of these assumptions in order to increase reproducibility [139]. This solution seems feasible as the treatment of missing values is a part of the

statistical analysis, and reporting them is in alignment with Reproducibility Type A.

3.1.5 Multiplicity

Multiplicity [64] or failure to adjust for multiplicities [19] can also lead to lower reproducibility. Multiplicity occurs when several statistical inferences are considered simultaneously, this often involves using multiple statistical tests. According to Bretz and Westfall [29], ignoring multiplicity in any stage of drug development may cause a lack of reproducibility, which they call *replicability*, at a later stage or after market approval. In a study conducted by Bretz and Westfall (25), simulations were performed with pairs of independent studies, the original and the replicate study. The only difference between the two studies was the sample size, where the original test study had sample size of 100, and the replicate study had a sample size of 1000. All other factors remained the same in both studies. They concluded that the effect sizes of the original study are not 'reproduced' in the replicate study: on average they are larger than effect sizes of the replicate study. This confirms that changing one aspect of the test, such as the effect size, can have an effect on the test conclusion.

3.1.6 Deliberate statistical malpractices

Intentional statistical malpractices are another cause of poor reproducibility. These include: removing 'outliers' and unfavourable data [19], trying out multiple models until one gets favourable results [19] (also called p-hacking [64, 108] or selective reporting [12, 64]), statistical overfitting [11], data dredging (analysing data in order to find any possible relationships between the data) [64] and hypothesizing after the results are known [64, 108] and questionable interim analysis (performing data analysis while still collecting data, and stopping when p-value is statistically significant), questionable inclusion of covariates [58] (adding covariates gradually to a regression model in order to find a significant effect), and questionable subgroup analyses [58] (reporting subgroup yielding the smallest p-value) [58]. Such malpractices often stem from the pressure to publish [12]. Clear documentation of all statistical processes, which links to Reproducibility Type A, allows an external scientist to check the analysis carried out and it increases the chance of spotting statistical malpractices.

Moreover, pre-registration of studies is conducive to transparency [28, 52, 105, 108, 141] and it prevents many malpractices. In pre-registration of studies, experimental designs and analytical plans are written down in a database before the experiment is performed. In clinical studies, pre-registration is mandatory and can be done through registries such as the International Standard Randomized Controlled Trial Number registry [85] and the International Clinical Trials Registry Platform [78].

3.2 More general insights

The majority of reasons for low reproducibility do not stem from wrong statistical analysis. Preference of publishing positive results, a lack of documentation of experiments, focus on exploratory studies rather than replication studies, and other non-statistical issues can lead to low reproducibility. This section will summarise these problems and outline suggestions for improvement offered in the literature.

3.2.1 Preference of publishing positive results

One of the reasons for low reproducibility is the pressure to publish [12]. This is exacerbated by the publication bias [19, 108, 142, 154], which refers to the preference of journals to publish positive results and reject negative results [57]. Similarly, negative or null results are also often not written up for publication. This leads to a high proportion of 'false positive' results.

Journals should strive to accept for publication articles with negative or null results [17, 19, 33]. Johnson et al. [88] argue that rather than based on statistical significance, articles should be accepted based on the quality of the experiment and the data, and the importance of the hypotheses tested. Removing the stigma associated with negative results, i.e. negative perception of negative results, has the potential to increase reproducibility [145]. However, even if journals allow the publication of negative results, it is questionable whether researchers will start writing up negative results simply because they are focused on positive results, and face many pressures which prevent them from writing negative results for journal publication. It is also questionable whether scientists would worry about the reproducibility of negative results. If not, false negatives would be more problematic for science than false positives because they would receive less attention and scrutiny.

3.2.2 Other problems in the publication system

Allison [1] emphasised that there is a lack of formal guidance for post-publication corrections. He pointed out that in science, a degree of self-correction is crucial. However, it is hard to achieve via publications. Once an article gets published, it is hard to address any errors. The US National Institute of Health (NIH) promoted that journals should be motivated to allocate more space for papers that point out errors in earlier work [33], which seems a feasible and forward-looking solution. Many journals have adopted this policy.

Other problems in the publication system that lead to lower reproducibility are fraudulent research [12, 108, 118], insufficient peer review, oversight and mentoring [12] and competition between laboratories leading to hastily written papers [57].

3.2.3 Documentation

Incomplete or bad reporting [57, 64, 144] or a lack of 'instruction material' for scientists who want to produce such reproducible research [52, 118] can also

lead to low reproducibility. The raw data, method description or code are not always available [12]. One of the reasons why scientists may not share their data and code is that it takes a lot of time to document the work and to clean up the data and code [97]. Alternatively, the researchers may not want to share documentation which includes work beyond the published results.

Iqbal et al. [84] carried out a systematic assessment of the biomedical literature, assessing transparency and reproducibility in a random sample of 441 articles in biomedical journals published between 2000 and 2014. They concluded that the biomedical literature lacks transparency; it is missing protocols, data, statements of conflict, funding information, and statements of novelty or replication. Similarly, Errington et al. [52] highlighted problems for replication of experiments: incomplete documentation, not enough information to repeat an experiment, descriptive or inferential statistics not provided, and insufficient detail about the experiments. In their reproducibility project in cancer biology, the team of scientists sought to repeat 193 experiments from 53 papers [52]. In order to do so, they had to modify 67% of the protocols, which were already peer-reviewed, and they were only able to implement 41% of those modifications [52]. Only 4 out of 193 experiments included data which were necessary for computing the effect size and for conducting power analysis [52]. Following difficulties faced during the design and conduct of the experiments, Errington et al. [52] were only able to carry out new experiments based on the original experiments for 50 experiments from 23 papers. Seibold et al. [133] attempted to reproduce analyses (aligned with Reproducibility Type A) of longitudinal data in 11 articles published in PLOS ONE, where authors consented and data were available. They were unable to reproduce results for three full articles and parts of two, concluding that reproducing results is difficult without provided code. Xiong and Cribben [153] attempted to reproduce 93 papers from prominent statistical journals using functional magnetic resonance imaging (fMRI) data. Without contacting the original authors, they successfully reproduced results for only 14 papers, all of which provided real fMRI data and executable code.

Many sources agree that careful documentation of all steps in an experiment is important for reproducible research [11, 13, 19, 59, 61, 108]. From the perspective of the five reproducibility types, clear documentation is important. This includes code, data and a clear description of the data and the analysis [30, 131]. Moreover, public access to all these documents is necessary so that other researchers can validate the analysis [130]. Haibe-Kains et al. [67] suggest that if data cannot be shared, an independent, highly trained investigator should verify the analysis. In fields like medical imaging AI, sharing data processing and training pipeline details is also essential. Berger [19] suggested establishing protocols for scientific investigations. Donoho [50] advised to create a single R script that generates all the results, figures and tables, for a particular paper. Solutions in terms of user-friendly software are not new; Schwab et al. [131] described ReDoc, a simple software system where authors deposit all the documentation, data, and code, that allows readers to

reproduce computational results from the articles. A more recent tool for a clear documentation of the statistical analysis is the R package *knitr* which creates a single document containing both the code and the documentation of the experiment, including visualisations [152]. Another solution is the use of a Jupyter Notebook. This is a web-based computational environment that shows the code for data analysis alongside text and visualisations [130]. All these solutions are compatible with the approach to reproducibility described in Reproducibility Type A, making it possible for another researcher to go through the data, code, and the method, and reanalyse the experiment. However, these solutions are not limited to Reproducibility Type A; they are useful for researchers who want to analyse the same data using a different analytical method (Reproducibility Type B) or for researchers who want to repeat the experiment (in line with Reproducibility Type C, D or E). A difficulty is that these solutions are time-consuming and require training. Nevertheless, the long-term benefits of these solutions are apparent, and they have already been implemented, for example, in computer sciences. A related question is what to do when irreproducibility is reported, in line with Reproducibility Type A or its stronger version.

The recommendations on documentation should not be limited to technical aspects, such as what software to use, but they also should include a discussion of what information should be included and in what depth. A detailed manual about journal reporting in quantitative research in psychology can be found in Appelbaum et al. [6]. The recommendations presented in this article can be also applied to other research fields. An earlier work includes Wilkinson et al. [150] who give a detailed and useful guide for practitioners in psychology on how to carry out appropriate statistical methods, devise a good experimental design, document the work well. This article does not limit itself to the field of psychology. Reporting guidelines for a broad spectrum of health research studies are given by the Enhancing the Quality and Transparency Of health Research (EQUATOR) network [54].

An example of efforts to encourage authors to share their data and code is the use of kite marks, introduced by the journal Biostatistics. Authors receive a 'C' kite mark for sharing code and a 'D' kite mark for sharing data. Additionally, authors may apply for a reproducibility review, and if reproducibility (aligned with the definition of Reproducibility Type A) is verified, they are awarded an 'R' kite mark [119].

Tiwari et al. [146] proposed a 'reproducibility scorecard' for publications to improve reproducibility. This scorecard asks 8 questions, two examples of these questions are: "Are the model codes deposited in a relevant open model database?" and "Are the mathematical expressions described in the manuscript or supplementary material?" [146]. Tiwari et al. suggested a 4 out of 8 cut-off in the reproducibility scorecard. This means that 'yes' needs to be answered to at least 4 question for there being a chance of reproducing the same results. Tiwari et al. [146] limited the scope of their paper to systems biology modelling but this idea could also be used in other scientific areas.

However, rather than using a cut-off-point, it might be better to report which 'reproducibility criteria' the publication satisfies. These reproducibility criteria could encompass all five reproducibility types.

In big data settings, keeping track of all steps in an experiment and data management becomes a challenge. There are many tools that make it possible for researchers to document their work, such as an open-source programming language, a cloud-based data repository, a programming interface and the previously mentioned Jupyter Notebook [135]. Moreover, following the FAIR (Findable, Accessible, Interoperable and Reproducible) principles [151] for data management can lead to higher reproducibility [130]. FAIR principles are guidelines that guide researchers on how to organise, describe, store and operate data in order to improve the reusability of data. See Wilkinson et al. [151] for a more elaborate description of these principles.

3.2.4 Cooperation

Within an organisation, better reproducibility, in alignment with all reproducibility types, can be achieved through collaboration of a team [105] and the inclusion of statisticians in research teams [19]. This is linked to the need for interdisciplinary teams on large scale projects [130] and for initiative to share common vocabulary. This would allow for more informed conclusions. Moreover, better mentoring and supervision, better teaching, more within-lab validation, incentives for diligent work, and more external-lab validation can improve reproducibility [12].

3.2.5 Focus on replication studies

Funding bodies also have a role to play as they can have an impact on reproducibility through grant distribution. There should be distinguishment between exploratory versus confirmatory studies [147]. Pusztai et al. [125] proposed that some of the existing funding from new-discovery oriented grants gets allocated to confirmatory and validation grants that could be used for verification of important published results. For example, Iorns et al. [83] presented a successful replication study in biology. The details of the experiment are omitted as these include biology related terminology. Iorns et al. [83] communicated with the original authors to receive further details of the study and they also performed additional analysis to collect more detailed data, including data of higher resolution than the original test scenario. This replicate study confirmed the conclusions of the original test scenario. However, the effects seen in the replicate study were lower than in the original study. To improve reproducibility, such efforts should receive recognition in the scientific community. However, such recognition might be hard to establish, as more emphasis is given to the discovery research and to the publication of novel findings. The Science Exchange network [132] established a support network for researchers who want to carry out replication studies in order to validate key experimental findings.

4 Statistical Reproducibility

Up to this point, this paper has categorised definitions of reproducibility, presented reasons for low reproducibility and suggestions on how to improve reproducibility, and discussed reproducibility within the context of preclinical research. This section provides a concise summary of debates on statistical reproducibility.

4.1 What is statistical reproducibility?

Similar to the term *reproducibility*, the term *statistical reproducibility*, *reproducibility probability* or *replication probability*, is not clearly defined. The first insights related to statistical reproducibility were provided by Goodman [63], who highlighted a misconception regarding the p-value. Goodman [63] questioned the claim that a small p-value improves the credibility of the test result and argued that the replication probability may be smaller than expected. Although Goodman used the term replication probability rather than reproducibility probability, his definition is similar to the definition of reproducibility adopted in this paper. Goodman [63] defined it as the probability of observing another statistically significant result in the same direction as the first one, if an experiment was repeated under identical conditions and with the same sample size, which is consistent with Reproducibility Type C. Senn [134] agreed with Goodman that the p-value and replication probability are different measures and that inconsistency between test results from individual studies may be expected. However, he disagreed with Goodman's claim that the p-value may overstate the evidence against the null hypothesis [63], both under the Frequentist and the Bayesian framework. According to Senn [134], under the Frequentist framework, p-value is the most rigorous possible type I error rate that could be considered and still lead to the rejection of the null hypothesis. Under the Bayesian framework, it could be argued that the p-value corresponds to a particular Bayesian posterior probabilities. Nevertheless, Senn [134] recognised that a link between the p-values and replication probability should be recognised. This article uses the term reproducibility probability (RP) instead of replication probability.

Miller [104] argued that there are two interpretations of the replication probability and that in both cases the probability is unknown. Miller called them the aggregate and the individual replication probability [104]. According to Miller, the former term refers to experiments being performed by different teams of researchers with varying conditions, which corresponds to Reproducibility Type E, whereas the latter term refers to experiments being carried out by a particular individual under exactly the same conditions, which corresponds to Reproducibility Type C and to Goodman's definition of statistical reproducibility. Miller discouraged researchers from attempting to estimate both types of replication probabilities, as, according to him, the initial data provide very little information about the RP in the follow-up experiment [104]. This is something we disagree with; we believe that the data from the original

test scenario can provide useful statistical insights; a statistician uses data for inference, hence, it contains further information.

Stodden [144] had a different approach to the use of the term *statistical reproducibility*. She described it as conception about how statistics affect the likelihood of a scientific result being reproducible and how they contribute to the study and the quantification of reproducibility [107]. Stodden also used this term to refer to the situation when flawed statistical analysis or experimental design leads to the failure to replicate the experiment [144]. The positive side of this definition is that it emphasises the importance of appropriate use of statistics in experiments. However, this definition generalises *statistical reproducibility* to any discussion regarding statistics and reproducibility, and it cannot be classified as any of the Reproducibility Types introduced in Section 2.

The debate on statistical reproducibility raises a variety of questions. In the proceedings of the workshop on statistical reproducibility by National Academies of Sciences, Engineering, and Medicine (NASEM)[107], one of the questions focused on what study designs and appropriate metrics can be used to quantify reproducibility of scientific findings. The proceedings of NASEM [107] mainly concentrated on the variability across studies, on how to assess this variability and on what degree of variability leads to worries about the lack of reproducibility. Indisputably, variability is an important factor in the statistical reproducibility debate. Lomax [107] explained that it is important to recognise which aspects of variation can and which cannot be controlled.

Exchangeability of random variables forms part of the variability discussion. De Finetti's Theorem [76] states that exchangeable observations are conditionally independent. It means that variables can be swapped around in the sequence, and following this their joint distribution does not change. Exchangeability can never be verified, but statisticians still make the assumption of exchangeability under the guidance of practitioners. In the reproducibility debate, it is important to ask whether or not one can assume exchangeability. This paper proposes that exchangeability could be assumed when the replicate experiment is carried out under the same conditions. This work assumes exchangeability in the nonparametric predictive inference (NPI) framework, as will be explained in Section 5.2.5. Thus, exchangeability can only be assumed for Reproducibility Type C. It is arguable whether exchangeability, or some extent of exchangeability, can also be assumed for Reproducibility Type D. Exchangeability can no longer be assumed for Reproducibility Type E, where the experiments are carried out under different conditions.

The proceedings of NASEM also discussed how statistics, in particular the choice of study design and analysis, can affect reproducibility of scientific results, and how reproducibility can be enhanced via structural and analytical approaches [107]. These questions address statistical causes of poor reproducibility and suggestions for improvements, both of these have been addressed in Section 3. However, these proceedings of the workshop [107] did not give

a summary of the existing metrics that are aimed at validating reproducibility and quantifying statistical reproducibility. This task will be pursued in Section 5.

Lastly, there is an important question: Within what framework should statistical reproducibility be assessed? BinHimd and Coolen [22, 41] considered reproducibility as a predictive problem and provided a frequentist approach, nonparametric predictive inference, to solve it. This paper adapts their approach to statistical reproducibility. Within the Bayesian framework, predictive inference has been discussed by Billheimer [21]. With a view to improve reproducibility, Billheimer [21] proposed predictive inference to predict observables. According to Billheimer [21], statistical modelling should predict observable quantities and events, based on the current data and other applicable information, rather than form inferential conclusions through hypothesis tests or estimation of parameters. Billheimer promoted that instead of focusing on unobservable parameters, attention should be centred on observable events. This view is in alignment with the approach to statistical reproducibility presented in this paper, however, this work suggests using NPI framework instead of Bayesian framework, as NPI does not make as many assumptions about the data as a Bayesian framework does.

4.2 The p-value and the statistical significance

Concerns regarding reproducibility of research results are interlinked with the ongoing debate about whether or not to use p-values [21]. In hypothesis testing, which is a method of statistical inference, p-values are used to make dichotomous decisions about whether to reject or fail to reject the null hypothesis. Depending on the p-value, test outcomes are labelled statistically significant or non-significant. The most commonly used threshold value in biomedical research for the p-value is 0.05. The p-value is the probability of obtaining the same or a more extreme value for the test statistic, under the assumption that the null hypothesis is correct. The American Statistician (TAS) [149] suggested abandoning the concept of *statistical significance* in scientific research. The grounds for this suggestions are that the concept of statistical significance is misinterpreted by many, that it can cause erroneous beliefs and poor decision making, and that it stops statistically insignificant results from being published. Furthermore, statistical significance does not imply truth, yet many researchers and bodies equate it with truth [149]. The editorial [149] stated that it is not enough to have directions, such as "Don't believe that an association or effect exists just because it was statistically significant", but that the p-values should not be dichotomised, i.e. test outcomes should not be labelled as statistically significant or non-significant, and the word statistically significant should not be used. TAS [149] suggested that rather than stating the p-value, its meaning should be described in words.

Fisher introduced p-values for the use at the exploratory stage to see if the experiment findings should be further investigated [111, 149]. They were not meant to lead to a dichotomous decision making rule, reject or not reject the

null hypothesis. The dichotomous nature of significance testing often leads to p-hacking, the misreporting of true effect sizes by researchers who want to publish and need significant results, as discussed in Section 3. Similarly to TAS, Amrhein et al. [4] argued that dependence on statistical significance threshold can be misleading, and they suggested not using statistical significance thresholds and reporting only precise p-values.

Amrhein et al. [4] argued that conclusions should not be based solely on whether the p-values are significant or non-significant. Other metrics, such as the effect size and power, are equally important in the statistical analysis of tests. Amrhein et al. [4] also addressed the problem of making over-confident claims based exclusively on p-value. Nuzzo [111] also highlighted that effect sizes are often ignored and the research focus is on whether there is an effect rather than on how big the effect is, while the latter question is often more important. Nuzzo [111] discussed the problem of overrelying on p-values in decision making. Halsey et al. [70] discouraged analysis based mostly on p-values because of "the wide sample-to-sample variability in the p-value" [70]. They proposed that the dichotomous yes-or-no decision should be reached using a variety of measures, in particular the effect size estimates and their 95% confidence intervals [70]. Colquhoun [34] raised the problem of high false discovery rate for p-value around 0.05 in significance testing. He illustrated this on tree diagrams for simple testing procedures and he explores it further via simulations; the false discovery rate is the ratio of the number of false positive results to the total number of positive test results. The false discovery rate is also high for tests with low statistical power.

Halsey [69] offered four alternative analysis approaches to augment or replace the p-value. First, he discussed the augmented p-value augmented with information about its variability. He suggested the p-value prediction interval as a possible tool to do so. The prediction interval characterises the uncertainty of the p-value of a future replicate study [43]. However, augmented p-values may cause confusion as their interpretation is not straightforward and their calculation relies on p-values. Secondly, Halsey suggested estimating effect sizes and their confidence intervals [69]. Thirdly, Halsey suggested the use of Bayes factors instead of p-values as more intuitive metrics for interpretation. Fourthly, Halsey suggested using the Akaike information criterion for model assessment. Being aware and using alternative methods for statistical analysis gives decision-makers more flexibility and more tools to make decisions. However, these tools do not replace p-values as they are different measures and communicate different messages.

Macnaughton [99] disagreed with the claims made by TAS [149], in particular, that abandoning statistical significance will lead to fewer false-positive errors in scientific research, and that it will enable easier replication of scientific research results [99]. According to Macnaughton [99], science and statistics aim at separating signal from noise in data and the p-value is a useful tool for determining whether the studied effect exists in the population [99]. Unfortunately, false-positives still persist in published research, i.e. a p-value which

implies that there is evidence for the alternative hypothesis, but in fact the null hypothesis is true. Macnaughton [99] argued that the critical threshold value provides a balance between the rates of false-negative errors, false-positive errors, and costs. Macnaughton acknowledged that some people may manipulate p-values (either because of a lack of knowledge or on purpose so that they can publish) and this is harmful to science. Macnaughton also pointed out that if researchers obtain a p-value above the critical value of a relevant journal and if they believe that the studied effect exists and it is important, then the researchers should create a more powerful research design and repeat the study to see if they can get convincing evidence for the existence of the effect. Ioannidis [82] also argued that significance is essential for activity in both science and non-science and that some filtering process is helpful to avoid drowning in noise.

Benjamin et al. [18] proposed to change the default p-value threshold for statistical significance from 0.05 to 0.005 for new discovery claims, to improve reproducibility and to label novel findings with p-values between 0.005 and 0.05 as suggestive evidence. Reproducibility was not explicitly defined by Benjamin et al. [18], it could be assumed that they referred to Reproducibility Types C, D or E, or a combination of these. Benjamin et al. [18] did not propose that this new threshold is used for decisions on whether to publish or not. Similarly, the proceedings of the workshop by NASEM [107] discussed the benefits of increasing the threshold for demonstrating statistical significance, through p-values or Bayes factors. It is doubtful whether this would increase reproducibility because p-values and reproducibility probability are different measures and there is inconsistency between test results from different studies, as has been discussed by Senn [134] and Goodman [63].

Leek and Peng [96] identified that there are more important discussions than the question of whether or not to use p-values. It is more important to focus on the improvement of researchers' education in statistics and evidence-based data analysis, teaching them to use statistical analysis correctly. We agree with their point of view; the p-value forms only a small part of the experiment, which follows experimental design, collection and handling of data, and summary statistics, and the problem of a lack of reproducibility in science cannot be solely blamed on the p-value.

5 Quantification of Statistical Reproducibility

Following the documentation of the study, it is essential to carefully check the study design, code, data analysis, and other relevant aspects, ensuring there are no errors in alignment with Reproducibility Type A. Reproducibility Type B, on the other hand, requires the use of the original study's data but with a different analytical method to determine whether the same conclusions are reached.

As discussed in Section 2, Reproducibility Types C, D, and E involve two key scenarios. The first scenario involves both the original and replicate

experiments, where the focus is on assessing whether the conclusions from the original study are replicated. The second scenario applies when only the original experiment has been conducted, and reproducibility is evaluated based on the available data and statistical analysis. While the first scenario has received considerable attention, with various methods developed to assess replication success, the second scenario has been less explored.

5.1 Quantifying Statistical Reproducibility with Replicate Studies

Errington et al. [53] described seven methods for the assessment of replication: (i) statistical significance: whether the p-value is less than 0.05 for the original positive results or whether the p-value is greater than 0.05 for the original non-significant results; (ii) original effect size in the replication 95% confidence interval; (iii) replication effect size in the original 95% confidence interval; (iv) replication effect size in the original 95% prediction interval; (v) meta-analysis combining original and replication effect sizes, leading to p-value less than 0.05 for the original positive results or to p-value greater than 0.05 for the original non-significant results; (vi) comparing whether the results had the same direction - in the evaluation of representative images the original and replicate outcome can have the same direction but a different statistical significance; (vii) comparing whether the replication effect size is less than or equal to the original effect size. A replicated study was assessed as successful if majority of the criteria (i) - (v) were satisfied (3 or more out of 5). The other two criteria, (vi) and (vii), were not included in this assessment of a successful replication, as they do not work for non-significant effects, i.e. cases when the null hypothesis is not rejected. The comparison of effect sizes showed that the median of effect sizes in the replication studies was 85% smaller than the median of effect sizes in the original experiments, and 92% of the replication effect sizes were smaller than the original effect sizes. Moreover, the original null effects were replicated for 80% of the original tests, whereas the positive findings were replicated for only 40% of the original tests.

Open Science Collaboration [112] evaluated reproducibility via the following criteria: significance and the same p-value cut-off point, effect sizes, subjective assessment of replication teams, and meta-analyses of the effect sizes. They concluded that while 97% of the original studies had a p-value below 0.05, only 36% of the replication studies had a p-value below 0.05.

Patil et al. [114] highlighted the problem that the p-value cut-off points do not account for variation [114]. Patil et al. [114] instead suggested the consideration of the effect expected in the replication study, examining the original effect. Patil et al. [114] defined the 95% prediction interval, which can be calculated via Equation (1).

$$\hat{r}_{\text{original}} \pm z_{0.975}\sqrt{\frac{1}{n_{\text{orig}}-3}+\frac{1}{n_{\text{rep}}-3}} \tag{1}$$

where $\hat{r}_{\text{original}}$ is the estimate of the correlation coefficient in the original study, n_{orig} and n_{rep} are the sample sizes in the original and the replication study, respectively; and $z_{0.975}$ is the 97.5% quantile of the Normal distribution [114].

Patil et al. [114] warned that a small sample size leads to a wide prediction interval, and thus, the assessment of the replication study could be non-informative for small sample sizes. Patil et al. [114] pointed out that in the Reproducibility Project: Psychology [112] by Open Science Collaboration, the replication study effect sizes were smaller than the original study effect sizes due to publication bias. This observation is in line with the observation made in Reproducibility Project: Cancer Biology [53].

5.1.1 High Throughput Experimentation

In high throughput experimentation (HTE), automated equipment is used to run a large number of tests simultaneously. Parallelisation is the key principle of HTE. High throughput experimentation is, for example, used in biological science laboratories to rapidly screen millions of samples. Assessment of reproducibility is a highly discussed topic in high throughput experimentation, where the replicate study often has a different sample size than the original study. In the replicate studies, only signals that were positive, interesting or significant in the original study are studied. Thus, the sample size and design in the replicate study differ from the original study. Moreover, scientists sometimes introduce test compounds in the replicate study that have similar characteristics to those selected as significant in the primary screen.

The metrics used to quantify reproducibility in HTE are the r-value [72], irreproducible discovery rate (IDR) [98] and maximum rank reproducibility (MaRR) [121], where the r-value is briefly described. A detailed discussion of these metrics is outside the scope of this paper. This short survey of available metrics aims to illustrate that this type of assessment has received considerable attention in the literature. Both Li et al. [98] and Philtron et al. [121] named Spearman's pairwise rank correlation as a commonly used method for assessing reproducibility in HTE. However, both sources agreed that it is not the most suitable method as Spearman's pairwise rank correlation's properties depend on how stringent the requirements for inclusion of genes are.

In the field of genomics, assessing whether findings from a primary study are replicated in a follow-up study has been explored [27, 73]. The terminology used is *replicability*, findings being replicated in another study. The studies conduct large-scale searches for rare true positives; one study is simultaneously examining many features. In the context of genome-wide association studies, the follow-up studies often examine only features that were identified as significant in the primary study.

For the test scenarios described above, Heller et al. [72] introduced the r-value as a metric to quantify the strength of replication [72], i.e. evidence against findings from a primary study being replicated in a follow-up study. A smaller r-value means stronger evidence in favour of replicability [137]. The Benjamini-Hochberg procedure can be used on the reported r-values to control

the false-discovery rate (FDR). Heller et al. [72] defined the FDR r-value for feature i as the lowest FDR level at which the finding is among the replicated ones. Heller et al. offered an online calculator of the r-value [74]. Meta-analysis is often used in genome-wide association studies. However, Heller et al. [72] argued that meta-analysis, pooling results across studies, is not an assessment of replicability, and they suggested adding the r-value to the statistical analysis.

5.1.2 Agreement indices

Assessment of whether a replicate study reached the same conclusions as the original study, in accordance to Reproducibility Type C and Type D, has also been assessed via agreement indices. Barnhart et al. [15] compared various agreement indices: the Pearson correlation coefficient, the mean-squared deviation, the intraclass correlation coefficient, the kappa statistic, the concordance correlation, the within-subject coefficient of variation, the coefficient of individual agreement, limits of agreement, coverage probability, and total deviation index. They identified the coverage probability as the preferred index for assessing agreement because it can be applied to both continuous and categorical data, and it is intuitive and easy to compute. These metrics are not described separately as they are not relevant to the rest of this paper.

5.1.3 Reproducibility from a Bayesian perspective

Reproducibility has been assessed from a Bayesian perspective [16, 71, 138]. For example, Held [71] introduced the sceptical p-value (p_S), a quantitative measure for *replication success*. The term *replication success* is not explicitly defined by Held [71]. We assume that it means that the findings of the original experiment are validated in the replicate experiment. The technique is suitable for tests which employ frequentist analysis. It considers p-values, sample and effect sizes of both the original and replication study. The method determines the largest confidence level $1 - p_S$ for the original confidence interval, at which replication success can be declared at level p_S [71]. The author preferred this method to meta-analysis because, according to him, exchangeability assumptions are not appropriate [71]. Held's argument is that, via the conduct of a replication study, researchers challenge the findings of the original study, which is an asymmetric task. The problem we encounter with this method is that the term *replication success* is not clearly defined, and the definition of the sceptical p-value involves this term.

5.2 Quantifying Statistical Reproducibility without Replicate Studies

The previous section discussed metrics assessing reproducibility in situations where both the original and the replicate experiments have been carried out. This section focuses on metrics which are calculated after only the original

study has been carried out. These metrics relate to the probability of getting the same decision in a follow-up study. This view of reproducibility is in alignment with Goodman's [63] definition of statistical reproducibility and Billheimer's [21] approach to predictive analysis. In the literature, less attention is paid to this approach to statistical reproducibility. This paper will highlight such an assessment.

5.2.1 Confusing reproducibility with other statistics

We have observed that some researchers interpret p-values, effect sizes, or confidence intervals as measures of reproducibility. For example, one view considers p-values or effect sizes as different ways to assess reproducibility. Another interpretation connects p-values to reproducibility probability, such that $p = 0.01$ corresponds to $\widehat{RP} \approx 0.73$ and $p = 0.0001$ corresponds to $\widehat{RP} \approx 0.97$. It is unclear how these $\widehat{RP}$ values are defined or calculated. Cumming [43] argued that confidence intervals contain information about replication. We disagree that p-values, effect sizes or confidence intervals are measures of reproducibility as they have a clear definition in statistics, and reproducibility or related terms are not part of those definitions; these are different concepts.

5.2.2 Peculiar metrics

In the literature, there are peculiar measures of reproducibility, such as Posavac's t_{rep} and Kileen's p_{rep}. Both metrics are linked to significance testing. According to Posavac [122], the probability of *statistically significant exact replication*, t_{rep}, can be calculated by subtracting the minimum difference for a statistically significant t-statistic from the difference in means observed in the initial study. Posavac presented a graphical method for calculating the probability of an exact replication being less than 0.05 for a two-tailed test [122]. However, it is not clear from the article how this would quantify the probability of the next experiment yielding the same conclusion. Because of the vagueness of the approach, it is unclear how to apply it in practice.

Killeen [93] argued that the probability of replicating an experiment can be estimated using the statistic p_{rep}. He defined p_{rep} as the replicate effect which is of the same sign as the effect found in an original experiment [93]. Killeen was motivated by the fact that the p-value is commonly misinterpreted. According to Killeen [95], p_{rep} can be estimated by viewing it as a function of the p-value (denoted by p), using the following formula:

$$p_{\text{rep}} \approx [1 + (\frac{p}{1-p})^{2/3}]^{-1} \tag{2}$$

Maraun and Gabriel [100] pointed out that Killeen's calculation and interpretation of p_{rep} and of the concept of reproducibility probability contain errors. Nevertheless, they credit Killeen's claim that replicability should play a key role in the assessment of empirical results [100]. Lecoutre et al. [95]

also recognised that p_{rep} is incorrectly defined, because of the confusion between 1-tailed and 2-tailed p-values. Another problem with Posavac's and Kileen's calculations of reproducibility is that both of these metrics are dependent on p-values, which are not measures of reproducibility, as explained in Section 5.2.1.

5.2.3 Estimated power approach

De Capitani and De Martini [46–48] adopted Goodman's definition of reproducibility probability, i.e. the probability of obtaining the same test result in a second, identical experiment. This corresponds to Reproducibility Type C, but they considered it as an estimation problem instead of a prediction problem.

De Capitani and De Martini [46–48] equated reproducibility probability to the true power of a statistical test. Their method is called the *estimated power approach* [136] and has been presented for the t-test, Wilcoxon rank-sum test [46] and they also developed reproducibility probability estimation for other nonparametric tests [47]. Shao and Chow [136] also advocated the estimated power approach. De Capitani and De Martini [47] argued that their methods provides useful information for evaluation of the stability of statistical test results. It is unclear what is the precise definition of the stability of test results and what is the benefit of the estimated power approach.

De Capitani and De Martini argued that many clinical trials cannot be done more than once or twice, mainly because of their budgets and time constraints [45]. However, for an experiment to be scientifically valid, it is often required that it is reproducible. De Capitani [45] argued that in such cases reproducibility of the experimental conclusions should be addressed as reproducibility of statistical significance [45] and this should be evaluated using reproducibility probability. We disagree with their statement as we believe the interest should be in reproducibility of conclusions rather than reproducibility of statistical significance and the two cannot be equated.

In their study, De Capitani and De Martini only focused on reproducibility when the null hypothesis was rejected. However, the approach proposed in this paper offers predictive inference for statistical reproducibility in both scenarios - when the null hypothesis is rejected and when it is not. This ensures a more comprehensive evaluation of reproducibility.

5.2.4 G × L adjusted p-value

It is hard to achieve standardisation in preclinical research and there has been a shift to embracing variability, as discussed in Section 6. In line with Reproducibility Type E, all conditions cannot be the same in the replicate experiment. Kafkafi et al. [90] described genotype-by-laboratory interaction (G × L) adjusted p-value, a metric that is aimed at accounting for variability in genotype influenced by environment. G × L adjusted p-value indicates the probability of replicating the result in additional laboratories [90].

The sensitivity of strains of mice (animals with identical genetics) to the environment is assessed by collecting results about different strains from different laboratories and determining how consistent is the phenotype, i.e. the set of observable characteristics. The G × L adjusted p-value is derived by estimating the interaction noise $\sigma^2_{\mathrm{G\times L}}$ from studies of a number of strains of mice in different laboratories. This provides information on the extent to which the p-value needs to be adjusted. For example, if the strain is very susceptible to the environment, the p-value adjustment is greater. The International Mouse Phenotyping Consortium (IMPC) strived to promote a public database of mutant lines of mice that could be available to all laboratories. The random lab model (RLM) adds the interaction noise $\sigma^2_{\mathrm{G\times L}}$ to the animal noise to create a base for determining phenotype differences [90]. The power is subsequently lowered and confidence interval of the estimated effect size is widened, accordingly, to ensure replicability. In theory, scientists could calculate G × L-adjusted p-values and confidence intervals. However, the method does not appear to be developed for a wider-use application. Kafkafi et al. [90] claim that reporting G × L-adjusted p-values and confidence intervals alongside the usual p-values and confidence intervals would increase replicability in preclinical research but they do not present reasons.

This approach seems, at first sight, appealing, as the calculation of the G × L adjusted p-value takes into account results from a variety of laboratories. However, even the G × L adjusted p-value is susceptible to errors. The feasibility of this method is related to the question whether it is possible to accurately estimate the G × L variability and if it is reasonable to trust this estimate. The variability in animal testing is complex, it does not only depend on the mouse batch and a particular laboratory, but also on the person who runs the experiment, the time of the day or the year, and the environment conditions, as will be discussed in Section 6.

5.2.5 Nonparametric Predictive Inference (NPI)

The quantification of statistical reproducibility, particularly in cases where only the original test scenario has been carried out, has received less attention. Some metrics related to this scenario were mentioned earlier; however, they primarily focus on the analysis process or environmental factors, rather than the data itself. The data aspect is critical, as it provides valuable insights into the variability that can influence reproducibility.

This authors suggest viewing statistical reproducibility as a prediction problem, focusing on the variability introduced by data and statistical methods. This perspective offers a distinct approach, differing from other factors that influence reproducibility. Rather than relying on classical frequentist methods, which are less suited for predictive problems, we propose the Nonparametric Predictive Inference (NPI) framework. NPI is designed for making inferences about future observations, making it well-suited for assessing reproducibility. It offers a frequentist alternative to Bayesian approaches, such as Billheimer's [21].

The NPI framework assumes that future observations are exchangeable with the observed data, meaning they are equally likely to fall into any of the intervals created by the ordered data. This allows for predictions without assuming the exact location of future data within these intervals. Uncertainty is quantified through lower and upper probabilities, based on all possible orderings of future observations among the observed data. For further details, refer to [36, 38, 39].

In the context of reproducibility, statistical reproducibility is defined as the probability of obtaining the same test outcome when the test is repeated under identical conditions. This aligns with Reproducibility Type C and is central to the NPI approach. In NPI, after performing a hypothesis test on the original sample of size n, we determine whether to reject H_0 or not based on the value of the test statistic. Next, we predict a future sample of size n, where all orderings of the n future observations among the n actual data observations are equally likely. We then determine whether H_0 is certainly rejected, possibly rejected, or possibly not rejected for each ordering of the future observations. We count all orderings for which the conclusion is certainly the same as for the actual test for the lower reproducibility probability. For the upper reproducibility probability, we include the 'possibly' orderings where the conclusion is the same as for the actual test.

For large sample sizes, computing exact lower and upper reproducibility probabilities can be challenging due to the exponential growth in the number of orderings. To address this, reproducibility probabilities can be estimated using methods like the NPI bootstrap method (NPI-B) [41] and the sampling of orderings method [42, 101, 102]. These methods provide estimates of the lower and upper reproducibility probabilities and offer flexibility for various applications beyond hypothesis testing, such as estimating population characteristics from randomized response data [3].

In practical research, at each stage of the process, decision-makers must decide whether to proceed with further study or repeat the test. The NPI reproducibility probability offers a valuable metric to support this decision-making process. To obtain the NPI reproducibility probability, the full data set is required. Different data with the same test statistic value can lead to different reproducibility values, highlighting the importance of using complete data for accurate assessments. However, the NPI reproducibility probability does not imply that the test outcome is "right" or "wrong." For that, traditional aspects of hypothesis testing, such as significance levels, power, and other related post-data metrics, remain essential. Therefore, NPI reproducibility should be considered alongside other statistical methods, including power analysis, effect size (ES), and p-value assessments, to provide a comprehensive understanding of the test's robustness.

Extensive research on NPI for reproducibility has contributed significantly to the field [2, 22, 37, 40, 41]. NPI has been used to study reproducibility for various statistical tests, including nonparametric tests, likelihood ratio tests, and tests for population quantiles.

6 Reproducibility in Preclinical Research: A Case Study

This section will address some of the issues regarding reproducibility of studies that specifically relate to preclinical in-vivo research, i.e. research carried out on animals, typically rodents. Preclinical research mostly focuses on the actual replication of an experiment in accordance to Reproducibility Type E, as due to the inevitable variations between experiments, it is impossible to have exactly the same conditions in two separate experiments. Arguably, this is impossible in any area. Quantifying reproducibility, in situations when only the original experiment has been carried out, has not received much attention in preclinical research.

6.1 Ethical issues

Animals are a fundamental part of preclinical research and the majority of the discussion on reproducibility in preclinical research is linked to them. Due to ethical issues, sample sizes in animal studies are small. Thus, poor reproducibility may be to some extent unavoidable. On the other hand, a follow-up study, which assesses reproducibility, increases the number of animals needed [127]. The 3Rs principles [55] provide guidance for researchers on how to responsibly conduct experiments in animal research. The 3Rs stand for replace - animals by non-sentient animals whenever possible; reduce - the number of animals; and refine - improve animal well-being. The 'reduce' principle is the most relevant one in the discussion on reproducibility herein and there is, arguably, a need for a move from the traditional focus on reducing the number of animals per experiment solely to a more integrated approach which also considers validity, robustness and reproducibility of experiments. The 'replace' and 'refine' principles are indirectly linked to the reproducibility debate: the more a researcher adheres to these principles, the more ethical ground there will be to repeat the experiment or to use a larger sample size. For the 'reduce' principle, an important question arises: Is it possible to improve reproducibility using smaller sample sizes, thus reducing the number of animal, assuming the experiment is set up optimally?

6.2 Challenges of using animal in research

Small sample size, linked to the ethical concerns, as well as to financial and practical reasons, is only one of the challenges a researcher faces when working with animals in preclinical research. The involvement of animals adds additional uncontrollable variability. Animals are very perceptive to small environmental changes, such as light and noise, and this can have an impact on the experiment.

Apart from the variations related to animal use, experiments may face the problem of inevitable variations, such as time lag, variation of apparatus and material [142]. Similarly, variability of standard reagents [12] can affect the

experimental outcomes. Slightly changing the experimental procedure or using different laboratories, or different animal strains are some of the reasons for low reproducibility of experiment [33]. Here strain stands for a group of animals that are genetically the same.

Stevens [143] named other reasons, with focus on animal use in comparative psychology: There is often repeated testing on more animals that are more expensive than rodents, such as parrots or primates. Also people may have more objections to testing on more intelligent animals. Therefore, as much data as possible are collected during one experiment. This exploratory data analysis may lead to data fishing. Furthermore, there is often limited species coverage and species are often substituted in a replicate study.

6.3 Recommendations offered in literature

Reynolds [126] pointed out the lack of adequate statistical training in preclinical research and he advocated training in statistics for researchers, specific to preclinical research. According to Reynolds [126], researchers should be taught to create the statistical design and carry out data sampling, before analysing the data and making inferences. The importance of statistical training has already been discussed in Section 3. However, not much attention has been paid to the details of such statistical training, possibly because a lot of the literature has been written by non-statisticians. It would be desirable to discuss in greater depth the methods that should be taught, the level of understanding of the methods that researchers should acquire, and the guidance on when a non-statistician should consult a statistician.

Spanagel [141] recommended a variety of measures that can be incorporated into the planning and design of an experiment in order to improve reproducibility: Prior to a new study, researchers could consider conducting a systematic review or potential meta-analyses of existing related studies, conduct a power analysis, pre-register experimental study protocols, as discussed in Section 3, and consider carrying out multi-centre preclinical studies. In the context of research on psychiatric disorders, Spanagel [141] advised researchers to consider using animal models that satisfy two psychiatric diagnostic classification systems, which are based on observations from clinical research [141], and it is important that the preclinical study reflects those. It is also advisable not to overcomplicate statistical analysis and to use only the methodology that the researcher has a good understanding of [141]. Richter [127] argued that the risk of bias could be prevented by random treatment allocation, blind administration of the treatment, and blind assessment of outcome. According to Richter [127], this could eliminate aspects of the experiment which lead to misleading results. However, it is arguable whether randomisation is preferable to carefully balancing an experiment with known factors.

Regarding the documentation of an experiment, diligently following ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines [7] improves reporting standards in animal testing [127] and thus makes replication of the experiments easier. ARRIVE guidelines provide directions on

reporting of ten essential items: study design, sample size, inclusion and exclusion criteria, randomisation, blinding, outcome measure, statistical methods, experimental animals, experimental procedures, and results. Moreover, reproducibility can be improved by making raw data available in accordance to FAIR principles [151] and by publishing negative findings [141], both recommendations have already been discussed in Section 3.

In preclinical research, discussions of reproducibility typically focus on adhering to good statistical practices and embracing the inherent variability introduced by the use of animals. NPI reproducibility research, however, does not concern itself with deviations stemming from the fact that animal testing is never conducted under identical conditions (e.g., mice may have slightly different properties, and new experiments are often carried out in different laboratories). Instead, it is solely concerned with the reproducibility of statistical tests based on the original test scenario data, including the description of the data and the statistical analysis. While much of the literature on reproducibility centres around whether an experiment can be reproduced under similar circumstances, this is irrelevant for NPI reproducibility because it does not involve conducting a second experiment. Simkus et al. [140] presented NPI-RB for a pairwise tests application in preclinical research. They explored whether there is any relationship between reproducibility and p-values or effect sizes. The initial findings showed that there is a trend that test statistics close to the test threshold are likely going to lead to lower reproducibility. They also explored the reproducibility of the final decision when multiple pairwise comparisons are carried out. This aspect of reproducibility has not been addressed elsewhere in the literature. It was shown that statistical reproducibility for the final decision is notably lower than reproducibility for separate pairwise comparisons.

6.4 Heterogenisation – embracing variability

In alignment with Reproducibility Type E, there is a body of literature suggesting that systematic heterogenisation rather than standardisation improves reproducibility in preclinical research [25, 91, 92, 127, 128, 147]. This literature focuses on experiments carried out on mice. Richter [127] argued that perfect homogenisation decreases inter-individual variation within a study population to zero, which leads to statistically significant results that cannot be generalised to slightly different conditions. This is also called the *standardisation fallacy*. Standardisation does not account for animals being responsive to the environment, also known as phenotypic plasticity [91]. This biological variation caused by phenotypic plasticity differs from random noise [147]. In preclinical research, it has been suggested to embrace variability through systematic heterogenisation in order to improve reproducibility [91].

Examples of heterogenisation named in the literature are using mice of diverse characteristics, such as mice of different age, sex and body weight, [129]; using different inbred strains of mice [147]; co-housing individuals of different strains of mice [127]; varying the housing conditions of mice [147];

varying husbandry and test procedures [129]; and carrying out the experiment on mice at different times [25] or in multiple laboratories [148]. For example, Bodden et al. [25] presented a study where systematic heterogenisation, adding variability, via carrying the experiment on mice at different times of the day improves reproducibility (Type E).

A possible tool for heterogenisation is the use of randomised block designs for the experiments. This can include using time or a batch as blocking factors [56, 92]. The latter is called the multi-batch design where the experiments are split into small batches of animals which are tested at different times. These 'mini-experiments' are then brought together in the statistical analysis. Karp et al. [92] showed how multi-batch design improves reproducibility in a syngeneic tumour case study. For the multi-batch design, they explored the following statistical analyses: meta-analysis, a fixed effect regression approach, a random effect regression approach and a pooled approach [92]. A pooled approach was not recommended for the statistical analysis as it ignores batch information. Meta-analysis and random effect regression were recommended by the authors for the analyses of multi-batch design experiments [92].

Embracing variability also addresses a problem that is interlinked with reproducibility: there is a high failure rate in translating research from preclinical to clinical studies [127]. Translating research means that conclusions about a new treatment reached in the preclinical stage of the drug development are validated in clinical research [127]. In a pharmaceutical context, it is desirable that the conclusions of a study remain the same even if the circumstances change, in order to increase the chance of a successful translation of the findings from preclinical to clinical studies, as the end goal of pharmaceutical research is to provide a new treatment. Thus, in the long-term, the focus on improving and quantifying reproducibility can also positively impact translating research from preclinical to clinical studies and, consequently, improve the efficiency of the drug development process.

7 Concluding remarks

The paper provided a comprehensive literature review on reproducibility, discussing the main debates and highlighting the lack of a universally accepted definition. Various definitions and related terms available in the literature are classified into five types. It was shown that sometimes different definitions are used for the same term and sometimes the same definition is used for different terms; some definitions are not clear; and often the term reproducibility is used without being explicitly defined.

Reasons for low reproducibility and suggestions for improving reproducibility offered in the literature were outlined. Many of the solutions simply entail adhering to good scientific practice and using appropriate statistical, experimental, and documentation methods, as well as fostering collaboration among different stakeholders.

Statistical reproducibility has also been a key topic of debate. Similar to the concept of reproducibility, statistical reproducibility is not a clearly defined term. Goodman [63] defined reproducibility as the probability of observing another statistically significant result in the same direction as the first one, assuming identical conditions and sample size.

Statistical discussions of reproducibility have focused on the variability across studies and how to control this variability. Important questions remain, such as whether the assumption of exchangeability is important for quantifying reproducibility and what framework should be used to assess it. Related to the reproducibility debate has been the ongoing discourse on the use of p-values. While there are many issues associated with p-values, there is currently no clear alternative that can be widely adopted by researchers.

The paper also reviewed metrics used to assess reproducibility when both the original and replicate experiments are conducted, but noted that less attention has been paid to quantifying reproducibility when only the original experiment is performed. A gap in the current debate is the lack of a clear understanding of what the original study data can reveal about reproducibility. This paper proposed treating reproducibility as a predictive problem, which can be addressed through frameworks such as Nonparametric Predictive Inference (NPI), a method that offers a way to quantify reproducibility using available data.

Finally, we briefly discussed reproducibility challenges in preclinical research, focusing on ethical concerns and offering possible solutions. A key insight is the shift from striving for homogeneity to embracing variability in preclinical research. An important question is what should a decision-maker do when reproducibility is low? A statistician would most likely advise that in such cases an experiment should be re-run, possibly with larger sample sizes. However, there are often ethical and financial constraints that make the replication of the experiment difficult. It is of future research interest to present an action plan for cases where reproducibility is low.

8 Competing interests

No competing interest is declared.

9 Acknowledgments

This work was performed under the EPSRC CASE PhD studentship with grant reference number EP/M507854/1. Furthermore, the authors gratefully acknowledge support from AstraZeneca for providing their guidance, and for further contribution to the studentship. The authors thank Natasha A. Karp for her valuable contributions to the discussion on reproducibility, particularly in preclinical research.

References

[1] Allison, D. D., Brown, A. W., George, B. J. and Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. *Nature*, 530, 27–29.

[2] Alqifari, H. N. (2017). Nonparametric predictive inference for future order statistics. *A thesis*, University of Durham. https://npi-statistics.com/pdfs/theses/HA17.pdf [Accessed: 29 June 2022].

[3] Alghamdi, F. M. (2022). Reproducibility of Statistical Inference Based on Randomised Response Data *A thesis*, University of Durham. http://etheses.dur.ac.uk/14783 [Accessed: 06 Sep 2023].

[4] Amrhein, V., Greenland, S. and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307.

[5] Anderson, J. A., Eijkholt, M. and Illes, J. (2014). Ethical reproducibility: towards transparent reporting in biomedical research. *Nature Methods*, 10, 843–845.

[6] Appelbaum, M., Kline, R.B., Nezu, A. M., Cooper, H., Mayo-Wilson, E. and Rao, S. M. (2018). Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, 73, 3–25.

[7] ARRIVE. (2023). ARRIVE guidelines. https://arriveguidelines.org/arrive-guidelines [Accessed: 8 August 2022].

[8] Assam, P. N. Mintiens, K., Knapen, K., Van der Stede, Y. and Molenberghs, G. (2010). Estimating precision, repeatability, and reproducibility from Gaussian and non-Gaussian data: A mixed models approach. *Journal of Applied Statistics*, 37, 1729–1747.

[9] Atmanspacher, H. and Maasen, S. (Eds.). (2016). *Reproducibility: Principles, Problems, Practices, and Prospects*, Wiley.

[10] Augustin, T., Coolen, F. P. A., de Cooman, G. andTroffaes, M. C. M. (2014). Introduction to Imprecise Probabilities. Wiley, Somerset.

[11] Bailey, D. H., Borwein, J. M. and Stodden, V. (2016) Facilitating reproducibility in scientific computing: principles and practice. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 141–167, Wiley.

[12] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454.

[13] Baker, M. (2016). Quality time: it may not be sexy, but quality assurance is becoming a crucial part of lab life. *Nature*, 529, 456–458.

[14] Barba, L. A. (2018). Terminologies for reproducible research. George Washington University. https://arxiv.org/pdf/1802.03311.pdf [Accessed: 17 May 2022].

[15] Barnhart, H. X., Yow, E., Crowley, A. L. et al. (2016). Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Statistical Methods in Medical Research*, 25, 2939–2958.

[16] Bayarri, M. J. and Mayoral, A. M. (2002). Bayesian design of "successful" replications. *The American Statistician*, 56, 207–214.

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

[18] Benjamin, D. J., Berger, J. O., Johannesson, M. et al. (2018) Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.

[19] Berger, J. (2012). Reproducibility of science: p-values and multiplicity. *Duke University, 8^th International Purdue Symposium on Statistics June 21, 2012.* http://www.stat.purdue.edu/symp2012/docs/Purdue_Symposium_2012_Jim_Berger_Slides.pdf [Accessed 7 May 2019].

[20] Bergman, R. G. and Danheiser, R. L. (2016). Reproducibility in Chemical Research. *Angewandte Chemie International Edition*, Editorial, 55, 12548–12549.

[21] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73, 291–295.

[22] BinHimd, S. (2014). Nonparametric predictive methods for bootstrap and test reproducibility. *A thesis*, University of Durham. https://npi-statistics.com/pdfs/theses/SB14.pdf [Accessed: 29 June 2022].

[23] Blackman, N. J.-M. (2004). Reproducibility of clinical data I: continuous outcomes. *Pharmaceutical Statistics*, 3, 99–108

[24] Blackman, N. J.-M. (2004). Reproducibility of clinical data II: categorical outcomes. *Pharmaceutical Statistics*, 3, 109–122.

[25] Bodden, C., von Kortzfleisch, V. T., Karwinkel, F., Kaiser, S., Sachser, N. and Richter, S. H. (2019). Heterogenising study samples across testing time improves reproducibility of behavioural data. *Scientific Reports*, 9.

[26] Bogomolov, M. and Heller, R. (2023). Replicability across multiple studies. *Statistical Science*, 38: 602–620.

[27] Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, 108, 1480–1492.

[28] Botvinik-Nezer, R. and Wager, T. D. (2022). Reproducibility in neuroimaging analysis: challenges and solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

[29] Bretz, F. and Westfall, P. H. (2014). Multiplicity and replicability: two sides of the same coin. *Pharmaceutical Statistics*, 13, 343–344.

[30] Buckheit, J. B. and Donoho, D. L. (1995). WaveLab and Reproducible Research. Technical report, Standford, CA.

[31] Carriquiry, A. L., Daniels, M. J and Reid, N. (2023). Editorial: special issue on reproducibility and replicability. *Statistical Science*, 38: 525–526.

[32] Collberg, C., Proebsting, T., Moraila, G. et al. (2014). Measuring reproducibility in computer systems research. Technical report. Department of Computational Science, University Arizona, Tucson. http://reproducibility.cs.arizona.edu/tr.pdf [Accessed 27 May 2022].

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

[34] Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*, 1:140216.

[35] Coolen, F. P. A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15, 21–47.

[36] Coolen, F. P. A. (2011). Nonparametric predictive inference. In: *International Encyclopedia of Statistical Science*, ed. Miodrag Lovric, 968–970. Springer, Berlin.

[37] Coolen, F. P. A., Coolen-Maturi, T. and Alqifari, H. N. (2018). Nonparametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, 47, 2527–2548.

[38] Coolen, F. P. A. and Alqifari, H. N. (2017). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *Journal of Statistical Theory and Practice*, 8, 591–618.

[39] Coolen, F. P. A. and Augustin, T. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124, 251–272.

[40] Coolen, F. P. A. and BinHimd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, 8, 591–618.

[41] Coolen, F. P. A. and and BinHimd, S. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov-Smirnov test. *Journal of Statistical Theory and Practice*, 14:26.

[42] Coolen, F. P. A. and Marques, F. J. (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of statistical theory and practice*, 14, 62.

[43] Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286–300.

[44] Cambridge English Dictionary. (2022). data Meaning [online]. Available at: http://dictionary.cambridge.org/dictionary/english/data [Accessed 25 May 2022].

[45] De Capitani, L. (2013). An introduction to RP-Testing. *Epidemiology Biostatistics and Public Health*, 10, 1–16.

[46] De Capitani, L. and De Martini, D. (2013). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistics Computation and Simulation*, 85, 468–493.

[47] De Capitani, L. and De Martini, D. (2016). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy*, 18, 1–17.

[48] De Martini, D. (2008). Reproducibility probability estimation for testing statistical hypotheses. *Statistics and Probability Letters*, 78, 1056–1061.

[49] Dictionary. (2023). conclusion Meaning [online]. Available at: https://www.dictionary.com/browse/conclusion [Accessed 14 February 2023].

[50] Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, 11, 385–388.

[51] Ehm, W. (2016). Reproducibility from the perspective of meta-analysis. In Harald Atmanspacher and Sabine Maasen (Eds.), *Reproducibility:*

Principles, Problems, Practices, and Prospects, 141–167, Wiley.

[52] Errington, T. M., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*,10:e67995.

[53] Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10: e71601.

[54] EQUATOR. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org/ [Accessed: 5 October 2023].

[55] Fenwick, N., Griffin, G. and Gauthier, C. (2009). The welfare of animals used in science: how the "Three Rs" ethic guides improvements. *Canadian Veterinary Journal*, 50, 523–530.

[56] Festing, M. F. W. (2014). Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *Institute for Laboratory Animal Research Journal*, 55, 472–476.

[57] Folkers, G. and Baier, S. (2016). A continuum of reproducible research in drug development. In Atmanspacher, H. and Maasen, S. (Eds.) *Reproducibility: Principles, Problems, Practices, and Prospects*, 141–167, Wiley.

[58] Freuli, F., Held, L. and Heyard, R. (2023). Replication success under questionable research practices – a simulation study. *Statistical Science*, 38: 621–639.

[59] Gamble, C., Krishan, A., Stocken, D. et al. (2017). Guidelines for the content of statistical analysis plans in clinical trials. *Journal of the American Medical Association*, 318, 2337–2343.

[60] Geisser, S. (1993). *Predictive inference: An introduction.* Chapman & Hall., New York.

[61] Gentleman, R. and Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16, 1–23.

[62] Gibb, B. C. (2014). Reproducibility. *Nature Chemistry*, 6, 653–654.

[63] Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, 11, 875–879.

[64] Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8, DOI:

10.1126/scitranslmed.aaf5027.

[65] Gosselin, R.-D. (2019). Guidelines on statistics for researchers using laboratory animals: the essentials. *Laboratory Animals*, 53, 28–42.

[66] Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 20200210.

[67] Haibe-Kains, B., Adam G. A. et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, 586: 7829.

[68] Haidich, A. B. (2010). Meta-analysis in medical research. *Hippokratia*, 14, 29–37.

[69] Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15, DOI:10.1098/rsbl.2019.0174.

[70] Halsey, L. G., Curran-Everett, D., Vowler, S. L. and Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, 12, 179–185.

[71] Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A*, 183, 431–448.

[72] Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111, 16262–16267.

[73] Heller, R. and Yekutieli, D. (2014) Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8, 481–498.

[74] Heller, R., Bogomolov, C., Benjamini, Y. ReplicabilityFDR http://www.math.tau.ac.il/~ruheller/App.html [Accessed: 6 September 2023].

[75] Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' Theorem for sampling from a population. *Journal of the American Statistical Association*, 63, 677–691.

[76] Hill, B. M. (1988). De Finetti's theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion). *Bayesian Statistics 3*, Bernardo, J. M. et al. (Eds.). Oxford University Press, 211–241.

[77] Hutton, H. N. and Williamson, P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics*, 49, 359–370.

[78] ICTRP. International Clinical Trials Registry Platform. https://www.who.int/clinical-trials-registry-platform [Accessed: 13 October 2023].

[79] ISO. (2017) Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation. https://www.iso.org/obp/ui/#iso:std:iso:21748:ed-2:v1:en [Accessed 2 November 2022].

[80] Ioannidis, J. P. A. (2005). Why most published research findings are false. *Public Library of Science Medicine*, 2, e124, 2005.

[81] Ioannidis, J. P. A. (2014). How to make more published research true. *Public Library of Science Medicine*, 11, e1001747.

[82] Ioannidis, J. P. A. (2019). The importance of predefined rules and prespecified statistical analyses: Do Not Abandon Significance. *Journal of the American Medical Association*, 321, 2067–2068.

[83] Iorns, E., Gunn, W., Erath, J., Rodriguez, A., Zhou, J. et al. (2014) Replication attempt: "effect of BMAP-28 antimicrobial peptides on leishmania major promastigote and amastigote growth: role of leishmanolysin in parasite survival". *Public Library of Science ONE*, 9: e114614.

[84] Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D. and Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *Public Library of Science Biology*, 14: e1002333.

[85] ISRCTN registry. International Standard Randomised Controlled Trial Number registry. https://www.isrctn.com/ [Accessed: 13 October 2023].

[86] Jarvis, M. F. and Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*, 37, 290–302.

[87] JCGM. (2012). International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 3rd edition.

[88] Johnson, V. E., Payne, R. D. et al. (2017). On the Reproducibility of Psychological Science. *Journal of the American Statistical Association*, 112: 1–10.

[89] Joint Committee for Guides in Metrology. (2008). Evaluation of measurement data – Guide to the expression of uncertainty in measurement. https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6 [Accessed 1 June 2022].

[90] Kafkafi, N., Golani, I. et al. (2017). Addressing reproducibility in singlelaboratory phenotyping experiments. *Nature Methods*, 14, 462–464.

[91] Karp, N. A. (2018). Reproducible preclinical research? Is embracing variability the answer? *Public Library of Science Biology*, 16, e2005413.

[92] Karp, N. A., Wilson, Z., Stalker, E., Mooney, L. et al. (2020). A multibatch design to deliver robust estimates of efficacy and reduce animal use - a syngeneic tumour case study. *Scientific Reports*, 10:6178.

[93] Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345–353.

[94] Lau, A. T. C. (2009). What are repeatability and reproducibility. *ASTM Standardisation News*.

[95] Lecoutre, B., Lecoutre, M.-P. and Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use, and interpret them. *Psychological Methods*, 15, 158–171.

[96] Leek, J. T. and Peng, R. D. (2015). *P* values are just the tip of the iceberg. *Nature*, 520, 612.

[97] LeVeque, R. J., Mitchell, I. N. and Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Journal Computing in Science & Engineering*, 14, 13–17.

[98] Li, Q., Brown, J. B., Huang, H. and Bickel, P. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5, 1752–1779.

[99] Macnaughton, D. B. (2019). Two problematic promises on page one of the TAS Special Issue on Statistical Inference. MatStat Research Consulting Inc. https://matstat.com/macnaughton2019a.pdf [Accessed: 20 June 2022].

[100] Maraun, M. and Gabriel, S. (2010). Killeen's (2005) p_{rep} coefficient: logical and mathematical problems. *Psychological Methods*, 15, 182–191.

[101] Marques, F. J., Coolen, F. P. A. and Coolen-Maturi, T. (2019). Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. *Journal of Statistical Theory and Practice*, 13, 15.

[102] Marques, F. J., Coolen, F. P. A. and Coolen-Maturi, T. (2019). Approximations for the likelihood ratio statistic for hypothesis testing between two beta distributions. *Journal of Statistical Theory and Practice*, 13, 17.

[103] McAlinden, C., Khadka, J. and Pesudovs, K. (2011). Statistical methods for conducting agreement (comparison of clinical tests) and precision (repeatability or reproducibility) studies in optometry and ophthalmology. *Ophthalmic & Physiological Optics*, 31, 330–338.

[104] Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review: Theoretical and review articles*, 16, 617–640.

[105] Munafò, M. R., Nosek, B. A., Bishop, D. V. M. et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021.

[106] Nature. (2018). Challenges in irreproducible research [Special Issue]. https://www.nature.com/collections/prbfkwmwvz [Accessed 21 July 2022].

[107] National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*, Washington, DC: The National Academies Press. https://doi.org/10.17226/21915 [Accessed 28 April 2023].

[108] National Academies of Sciences, Engineering and Medicine. (2019). *Reproducibility and replicability in science*, Washington, D.C: The National Academies Press. https://doi.org/10.17226/25303 [Accessed 28 April 2023].

[109] Nosek, B. A., Hardwicke, T. E., Moshontz, H. et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–48.

[110] NSF. (2015). 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.* http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf [Accessed 5 June 2022].

[111] Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506, 150–152.

[112] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349. doi:10.1126/science.aac4716 [Accessed 20 May 2022].

[113] Parmigiani, G. (2023). Defining replicability of prediction rules. *Statistical Science*, 38: 543–556.

[114] Patil, P., Peng, R. D. and Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11, 539–544.

[115] Patil, P., Peng, R. D. and Leek, J. T. (2016). Supplement for "A statistical definition for reproducibility and replicability". https://www.biorxiv.org/content/biorxiv/suppl/2016/07/29/066803.DC1/066803-1.pdf [Accessed 20 May 2022].

[116] Patil, P., Peng, R. D. and Leek, J. T. (2019). Visual tool for defining reproducibility and replicability. *Nature Human Behaviour*, 3: 650–652.

[117] Peers, I. S., South, M. C., Ceuppens, P. R., Bright, J. D. and Pilling, E. (2014). Can you trust your animal study data? *Nature Reviews Drug discovery*, 13, 560.

[118] Peng, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics*, 3, 405–408.

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

[120] Peng, R. D., Dominici, F. and Zeger, S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*, 163, 783–789.

[121] Philtron, D., Lyu, Y., Li, Q. and Ghosh, D. (2018). Maximum rank reproducibility: a nonparametric approach to assessing reproducibility in replicate experiments. *Journal of the American Statistical Association*, 113, 1028–1039.

[122] Posavac, E. J. (2002). Using p-value to estimate the probability of a statistically significant replication. *Understanding Statistics*, 1, 101–112.

[123] Possolo, A. (2023). Tracking truth through measurement and the spyglass of statistics. *Statistical Science*, 38: 655–671.

[124] Prinz, F., Schlange, T. and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Drug discovery*, 10, 328–329.

[125] Pusztai, L., Hatzis, C. and Andre, F. (2013). Reproducibility of research and preclinical validation: problems and solutions. *Nature Reviews Clinical Oncology*, 10, 720–724.

[126] Reynolds, P. S. (2022). Between two stools: preclinical research, reproducibility, and statistical design of experiments. *BMC Research Notes*, 15, 73.

[127] Richter, S. H. (2017). Systematic heterogenization for better reproducibility in animal experimentation. *Laboratory Animals*, 46, 343–349.

[128] Richter, S. H., Garner, J. P. and Würbel, H. (2009). Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nature Methods*, 6, 257–261.

[129] Richter, S. H., Garner, J. P., Auer, C., Kunert, J. and Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. *Nature Methods*, 7, 167–168.

[130] Schaduangrat, N., Lampa, S., Simeon, S., Gleeson, M. P., Spjuth, O. and Nantasenamat, C. (2020). Towards reproducible computational drug discovery. *Journal of Cheminformatics*, 12, DOI: 10.1186/s13321-020-0408-x.

[131] Schwab, M., Karrenbach, M. and Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science & Engineering*, 2, 61–67.

[132] The Science Exchange Network. Validating key experimental results via independent replication. http://validation.scienceexchange.com/#/ [Accessed: 8 August 2022].

[133] Seibold, H., Czerny, S., Decke, S. et al. (2021). A computational reproducibility study of PLOS ONE articles featuring longitudinal data analyses. *PLoS One* , 16: e0251194.

[134] Senn, S. (2002). A comment on 'A comment on replication, p-values and evidence'. *Statistics in Medicine* (Letter to the editor), 21, 2437–2444.

[135] Arnaud Serret-Larmande, Jonathan R Kaltman and Paul Avillach. Streamlining statistical reproducibility: NHLBI ORCHID clinical trial results reproduction. *Journal of the American Medical Informatics Association Open*, 5, ooac001.

[136] Shao, J. and Chow, S.-C. (2002). Reproducibility probability in clinical trials. *Statistics in medicine*, 21, 1727–1742.

[137] Shenhav, L., Heller, R. and Benjamini, Y. (2015). Quantifying replicability in systematic reviews: the r-value. Cornell University, 2015. https://arxiv.org/pdf/1502.00088.pdf [Accessed: 28 April 2023].

[138] Shiffrin, R. and Chandramouli, S. (2016). Model selection, data distributions, and reproducibility. In Harald Atmanspacher and Sabine Maasen (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 141–167, Wiley.

[139] Sidi, Y. and Harel, O. (2018). The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209, 169–173.

[140] Simkus, A., Coolen, F. P. A., Coolen-Maturi, T., Karp, N. A. and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31, 673–688.

[141] Spanagel, R. (2022). Ten points to improve reproducibility and translation of animal research. *Frontier in Behavioral Neuroscience*, 16: 869511.

[142] Stahel, W. A. (2016). Statistical issues in reproducibility. In Harald Atmanspacher and Sabine Maasen (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 141–167, Wiley.

[143] Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, 8, 1–6.

[144] Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2, 1–19.

[145] Teixeira da Silva, J. A. (2015). Negative results: negative perceptions limit their potential for increasing reproducibility. *Journal of Negative Results in BioMedicine*, 14:12.

[146] Tiwari, K., Kananathan, S., Roberts, M. G. et al. (2021). Reproducibility in systems biology modelling. *Molecular Systems Biology*, 17:e9982.

[147] Voelkl, B., Altman, N. S., Forsman, A. et al. (2020). Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience*, 16, e2003693.

[148] Voelkl, B., Vogt, L. and Sena, E. S. and Würbel, H. (2018). Reproducibility of pre-clinical animal research improves with heterogeneity of study samples. *Public Library of Science Biology*, 16, e2003693.

[149] Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019). Moving to a world beyond $p < 0.05$. *The American Statistician*, 73, 1–19.

[150] Wilkinson, L. and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

[151] Wilkinson, M. D., Dumontier, M. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

[152] Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch and Roger D Peng (Eds.), *Implementing Reproducible Research.* CRC Press. osf.io/s9tya/ [Accessed: 26 May 2022].

[153] Xiong, X. and Cribben, I. (2023). The state of play of reproducibility in statistics: an empirical analysis. *The American Statistician*, 77: 115–126.

[154] Zwaan, R. A., Etz, A., Lucas, R. E. and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, e120, 1–61.