Reporting *In Vitro* Experiments Responsibly – the RIVER Recommendations

The RIVER working group

Abstract

Publications reporting *in vitro* experiments often lack basic details that prevent readers from repeating the experiments described or assessing whether the results are reliable. This prevents the experiments adding to the knowledge base and can lead to unnecessary subsequent studies. To address this, an international working group has been convened to develop reporting standards, drawing from its expertise in research funding and publishing, methodology and statistics, and research in academic, regulatory and industry settings. Here we present RIVER (Reporting *In Vitro* Experiments Responsibly), a set of six recommendations specifically tailored to reporting *in vitro* studies such that manuscripts describe the minimum information necessary for a reader to assess the methodological rigour and reliability of the study.

Introduction

In vitro experiments (involving, for example, cell cultures, organoids, microphysiological systems, or cell free models) make up a significant proportion of biological research and offer various insights for the study of many living systems and processes. Manuscripts describing *in vitro* studies need to contain sufficient information so that readers, editors and reviewers can evaluate the science presented and understand how the experiments were conducted. However, the quality of reporting of *in vitro* experiments is highly variable [1, 2]. Important information that would enable assessment of a study's reliability is frequently missing from manuscripts; this affects the ability of researchers and policy makers to decide whether to base future experiments, safety assessments or policy decisions on published results. Few resources exist that specifically describe which aspects of study design and conduct impact the reliability of an *in vitro* experiment, and therefore which details are important to report. This contrasts with other fields of biological science, in which reporting guidelines are widely recommended, frequently endorsed by journals and research funders, and increasingly commonly used. This manuscript defines a minimum standard for the reporting of *in vitro* experiments, with the aim of improving the reliability and transparency of published *in vitro* research, thereby increasing its value and usefulness to readers.

Evidence suggests that *in vitro* studies are often not reported with sufficient transparency for readers to appraise their methodological rigour and interpret the results. Two recent systematic reviews of drug discovery studies using cancer cell lines highlight the scale of the issue [3, 4]: the majority of papers analysed in these reviews lacked basic information about experimental procedures, such as the type of control used, how cell lines were identified, cell passage numbers or the concentration of key medium constituents. In one analysis, a quarter of the papers also lacked even the most rudimentary information on cell culture conditions, such as cell density, carbon dioxide concentration or temperature. Significant differences between the results of the studies analysed could not be explained based on the limited experimental information reported in the papers, which raises concerns about the reliability of these studies and their translational value. Most *in vitro* studies also lack information pertaining to the internal validity of the experiments, such as details about whether randomisation or blinding were used, the criteria used to exclude data, or how sample sizes were determined [4, 5]. Without clarity on the experimental approach and statistical methods used, a reader cannot evaluate the extent to which bias may have influenced a study's results, whether the analysis is appropriate, or how generalisable the findings may be.

Concerns about the lack of reproducibility in biological research have grown in recent years [6-8]. A 2016 survey by *Nature* showed that over 75% of biological researchers (including those working *in vitro*) have tried and failed to reproduce an experiment, with underreported methods and selective reporting of positive results identified as common factors [9]. The Reproducibility Project: Cancer Biology (a long-term project attempting to replicate the experiments found in 53 high-impact cancer biology papers [10]) found that none of the 193 experimental methods under investigation could be replicated using only the information in the original manuscripts. Every method investigated required some degree of clarification from the original authors, with more than 50% requiring clarifications classified as moderate, strong or extreme [11]. The project was eventually only able to replicate the methodology of 50 experiments (35 of which were *in vitro*) [12]. These

data demonstrate not only the regularity with which crucial details are absent from published studies, but also the direct consequences this has on the ability to critically assess the reliability of published studies, and replicate published methods in practice.

Studies that cannot be relied upon or replicated represent a waste of time, financial and material resources, and – in cases where samples derive from *in vivo* sources – animals. Including all the important experimental information in publications is beneficial to readers, other researchers and stakeholders in biological research. It allows research funding bodies to be confident that the science they have supported can be used to inform future research and policy, and builds confidence in the results published in scientific journals. *In vitro* experiments can directly impact the use of animals in biological research, either because they use animal-derived samples and reagents, or because they may lead to subsequent animal experiments. *In vitro* models also represent important replacement opportunities for some animal studies, but for these models to be credible, and for the research community to have confidence in them, they need to be reported to the same standards expected of animal research, such as those laid out in the ARRIVE guidelines [13].

Reflecting the scientific community's concerns around the reporting of *in vitro* experiments, a number of ongoing initiatives address the quality of *in vitro* reporting to some extent [14, 15]. For example, the MDAR (Materials Design Analysis Reporting) initiative lays out a general framework for transparent reporting, applicable to all types of biological experiment, which is designed to raise minimum reporting expectations across the life sciences [16]. Guidance documents on Good Cell and Tissue Culture Practice, originally published by the EU Reference Laboratory for alternatives to animal testing (EURL ECVAM) in 2005 and updated in 2022 [17, 18] were developed for practical use in the laboratory to assure the reproducibility of *in vitro* research, and cover a number of key principles at a high level. A recent publication led by several US federal agencies also proposed a technical framework for incorporating measurement quality features into *in vitro* protocols [19]. Meanwhile, the OECD's Guidance Document on Good In Vitro Method Practices (GIVIMP) initiative focuses primarily on the acceptance and validation of *in vitro* data for regulatory purposes, laying out comprehensive guidelines for every element of laboratory procedure [20]. Finally, the International Society for Stem Cell Research have announced an initiative addressing best practices and reporting recommendations for stem cell research, due for release in the near future [21]. In recent years, major research funding organisations have also emphasised the importance of rigour and transparency in preclinical research, including by requiring researchers to provide information on experimental design and methodology when applying for funding, and when reporting experiments. Examples include the National Institutes of Health (NIH) in the USA [22-24] and UK Research and Innovation (UKRI) funders in the UK [25]. These efforts have also been mirrored in some journals, which have established requirements for the reporting of *in vitro* studies to varying degrees.

The RIVER recommendations

To facilitate more transparent reporting of *in vitro* experiments, this manuscript proposes the Reporting *In Vitro* Experiments Responsibly (RIVER) recommendations, a concise set of reporting recommendations

specifically tailored to *in vitro* studies. These consist of a short list of the most important pieces of information to include in any paper describing *in vitro* experiments – the minimum information necessary for a reader to assess the methodological rigour and reliability of the study. This focus on reliability, and the prioritisation of a small number of recommendations, distinguishes RIVER from other, more holistic initiatives. The primary purpose of RIVER is to encourage researchers to consider the factors that may influence the reliability of their results and report those factors transparently. Importantly, the recommendations are not designed to act as a tool for assessing reporting quality, or for evaluating the internal validity (risks of bias) of a study.

The RIVER recommendations are general enough to apply to any *in vitro* experiment, but are not comprehensive – they do not attempt to define all the information that should be included in an *in vitro* manuscript. The small number of recommendations is intended to make adoption of RIVER by researchers, journals and other members of the scientific community quick and straightforward. They are also not designed to exclude or downplay any other information relevant to specific types of experiments. The recommendations provide a reference point for authors, editors and reviewers, helping to ensure that manuscripts contain the minimum set of important details necessary for readers to understand the experiments described, and to assess their reliability. Adhering to the RIVER recommendations may also improve the reproducibility of published methods, however this cannot substitute for the sharing of a detailed protocol (see Recommendation 4).

The recommendations have been developed by a diverse, international working group, with expertise from across the scientific community, including researchers from academia, industry and government agencies, statisticians and methodologists, and representatives from journals and research funding organisations. The variety of organisations represented helps to ensure that the recommendations are appropriate to the needs of a wide range of users. Details of the process used to develop the recommendations can be found in Supplementary Information.

A detailed explanation accompanies each recommendation, to ensure that it is well understood by researchers and can be followed in practice. The explanations provide the rationale and supporting evidence behind each recommendation and the details that should be reported to satisfy them. The recommendations include guidance on reporting of experimental procedures, data handling and presentation, and concepts associated with ensuring internal validity (i.e. minimising risks of bias) for *in vitro* study designs. Consulting the explanations when planning experiments will promote the use of rigorous experimental methodology that can then be reported transparently.

RIVER includes six recommendations (shown in Table 1), classified into three themes: 1. experimental design; 2. experimental procedures and materials, and 3. data handling, accessibility and visualisation. These recommendations apply specifically to controlled experiments. While the recommendations are presented and discussed sequentially, they are interconnected and often need to be considered simultaneously.

Theme: Experimental design	
Item	**Recommendation**
1. Experimental unit	Define the experimental and biological units used in each experiment
2. Risks of bias	Report whether and how risks of bias were considered and addressed in the design of each experiment
Theme: Experimental procedures and materials	
Item	**Recommendation**
3. Experimental Model	Provide details of the model used in each experiment, including (if applicable) the identity, source and life stage of any cells used, information on microenvironmental (i.e. physical and biochemical) conditions, and quality control metrics
4. Experimental procedures	For each experiment, describe the experimental procedures in detail, including what was done, when and how often, and using which equipment and reagents
Theme: Data handling, accessibility and visualisation	
Item	**Recommendation**
5. Experimental groups and exclusions	Report all data obtained from all experimental groups (including controls) and justify any exclusions
6. Data availability and presentation	Share and present data transparently

Table 1: The RIVER recommendations

Recommendation 1: Experimental unit

Define the experimental and biological units used in each experiment

What is the experimental unit?

When designing an experiment, it is crucial to identify and report the specific entity on which the experiment is being carried out – the experimental unit. This information allows a reader to understand how the hypothesis is being tested, the design of each experiment, and the number of times the experimental procedures have been replicated, i.e. the sample size [26]. For this purpose, it is useful to understand the concept of the experimental unit, and distinguish between it and other types of biological entity.

The experimental unit is the entity that is randomly and independently assigned to different experimental groups (treatment conditions) during an experiment. Identifying the experimental unit is necessary to understand the sample size of an experiment, as the sample size of each group (*n*) is (providing certain conditions are fulfilled) equal to the number of experimental units in that group. For an entity to be considered an independent experimental unit, it is useful to consider three criteria [27-31]:

1. It should be possible to randomly and independently assign an experimental unit to any experimental group (see Recommendation 2 for a discussion of randomisation).
2. Individual experimental units should have interventions (e.g. drug treatments) applied to them independently of all other experimental units.
3. Experimental units should not influence each other, either within or between experimental groups.

For example, a simple cell culture experiment may be designed such that each well of a twelve-well cell culture plate (into which cells have previously been plated) is randomised to receive either a vehicle control or a drug treatment. That is, each well is randomised to either the control group or the drug treatment group. Following vehicle or drug treatment, the individual cells in each well are analysed by microscopy, with the whole experiment taking place on the same day. In this case, because the individual wells of the plate were randomised to different experimental groups, the interventions were provided to each well independently, and the wells could not influence each other, the experimental unit is the individual well. If an equal number of wells were randomised to each group, the sample size in this experiment is n=6 for the control group and n=6 for the drug treatment group. This assumes only a single control and single drug dose are used; in practice a rigorous experimental design may include positive as well as negative controls, and multiple drug dose levels.

Identifying the experimental unit

Identifying the experimental unit can sometimes be difficult. For this reason, it is useful to contrast the term with other related, but distinct, terms that can be used to describe entities in *in vitro* experiments. Examples of these include ‘biological unit of interest’ and ‘observational unit’:

- The biological unit of interest is the entity that a researcher wants to make an inference about. The purpose of an experiment is to test a hypothesis or to estimate a property regarding these biological units. The biological unit of interest could be an animal, a cell, a tissue, an organelle, or another kind of biological entity.
- The observational unit is the biological entity from which measurements are taken, which may or may not correspond to the experimental unit.

The three-way distinction between experimental, biological and observational units is important, because in many cases the three may be distinct from one another. In particular, the relationship between experimental units and biological units is complex, because an experimental unit may correspond to [30, 31]:

1. A single biological unit of interest,
2. A group of biological units,
3. A part of a biological unit, or
4. A sequence of observations on one biological unit.

In the previous example, although the experimental unit is the well, the observational unit (from which measurements are taken) is the individual cell, whilst the biological unit of interest may be the cells (or cell type), the organism from which they are derived, or a constituent of each cell (such as an organelle), depending on the hypothesis being tested. This can make it difficult to identify which unit corresponds with the sample size of an experiment, leading to problems with experimental design and the validity of results. Figure 1 shows examples of experimental units in different *in vitro* experiment types.

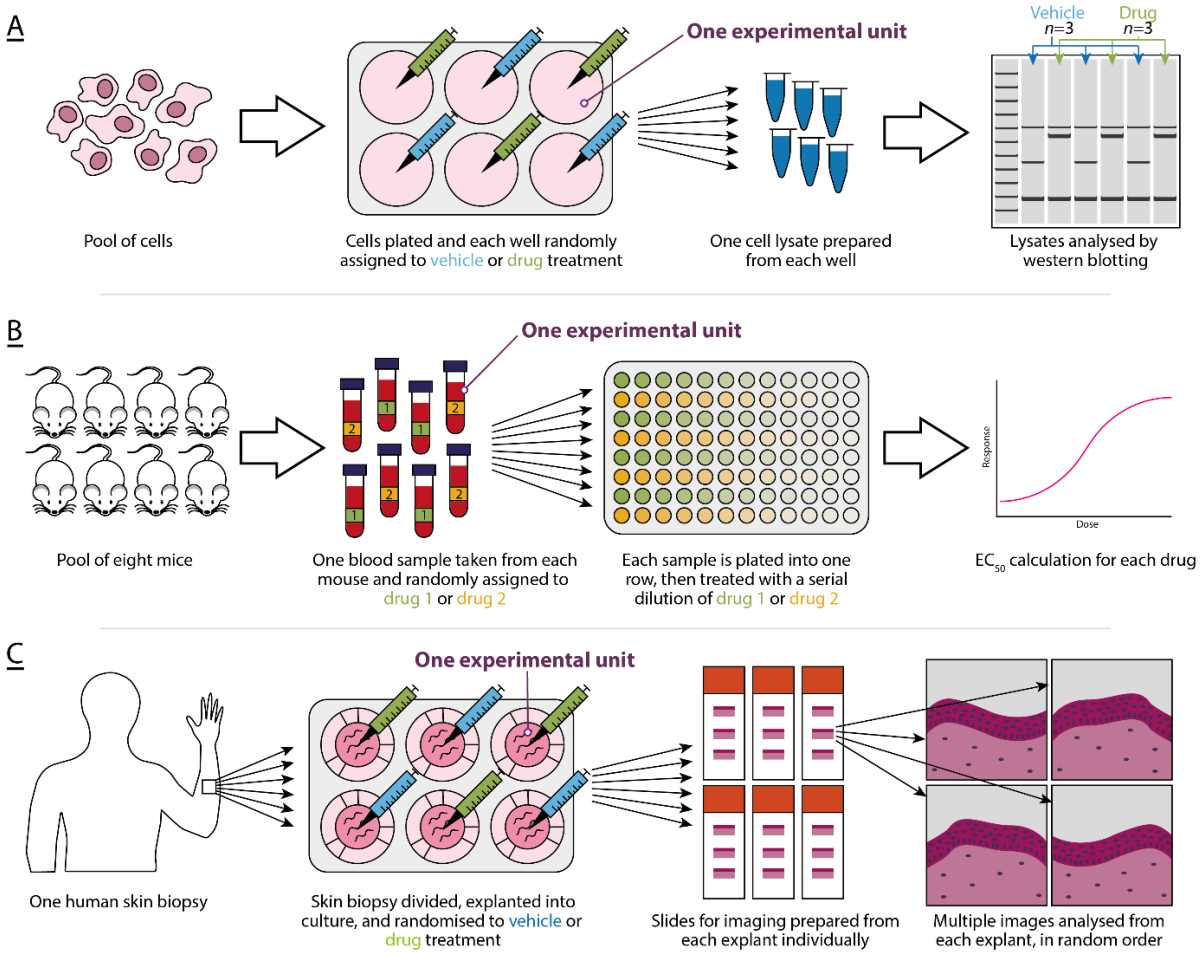

Fig 1. Identifying the experimental unit in *in vitro* experiments

(A) In this experiment, cells are seeded into the wells of cell culture plates, and each well is randomised to either a vehicle control or drug treatment group. After treatment, one cell lysate is made from the cells in each well. The lysates are then analysed by western blotting, with lysates from vehicle and drug groups alternated on the blot to account for any spatial effects (see Recommendation 2). In this case, because each well is randomised to an experimental group, the treatment is applied independently to each well, and the cells in each well cannot influence the cells in other wells, the experimental unit is each well. If the same lysate is run in multiple lanes on the western blot, these represent subsamples, or pseudoreplicates. Subsamples do not increase the sample size, even though they do increase the number of data points; results from subsamples can be averaged into one data point per experimental unit during statistical analysis. See 'the sample size, genuine replication and pseudoreplication' for more information.

(B) In this experiment, one blood sample is taken from each mouse. Each sample is randomised to receive one of two drugs and is then plated into one row of a 96-well microtitre plate (with samples assigned to the two drugs alternated to account for any differences that could arise from the position of the sample on the plate). Each row is treated with a serial dilution of one drug, and responses to each drug are used to calculate its half maximal effective concentration (EC_{50}). Each

blood sample is randomised to an experimental group, the treatment is applied independently to each row, and one measured response (the EC_{50}) is generated for each sample. In this case the experimental unit is the blood sample; the number of mice, rows of the microtitre plate and blood samples are all equivalent, and correspond to the sample size.

(C) In this experiment, one skin biopsy is taken from a single, healthy, human subject. This biopsy is divided into six samples, which are explanted and cultured in a cell culture plate. Each explant is randomised to either control or drug treatment conditions, after which multiple sections are taken from each sample for imaging analysis. In this case, because each explant is randomised to a treatment condition, the treatment is applied independently to each explant, and the explants in different wells cannot influence each other, the experimental unit is the explant. Where multiple images from each explant are analysed, these represent subsamples, or pseudoreplicates, as described in 1A.

When designing an experiment, it is useful to consider how the design impacts the applicability of any findings outside the context of the study (that is, the experiment's generalisability). Investigating a hypothesis in only one type of biological unit (such as one cell line, animal, strain or individual person) can limit the generalisability of a study's findings, as it may not be clear to what extent those results are applicable to different conditions. Generally, it is important to keep in mind that any inferences drawn from the results of an experiment can only apply to the population from which the experimental units are drawn (such as the cell line or strain of animal). Increasing the sample size of an experiment (i.e. increasing the number of experimental units) can strengthen support for the conclusions within that population, but may not increase the applicability of those conclusions to other populations (i.e. other cell lines, types of cells, strains or species of animal). In contrast, replicating experiments using diverse, independent biological units, such as multiple monoclonal cell lines, different strains of animal or groups of people can provide evidence to support the relevance of their findings more broadly. Reporting the biological unit for each experiment (in addition to the experimental unit) helps readers and researchers interpret the results of an experiment. Figure 2 shows an example of two different experimental designs for an *in vitro* experiment with the same sample size, and their impact on the generalisability of its results.

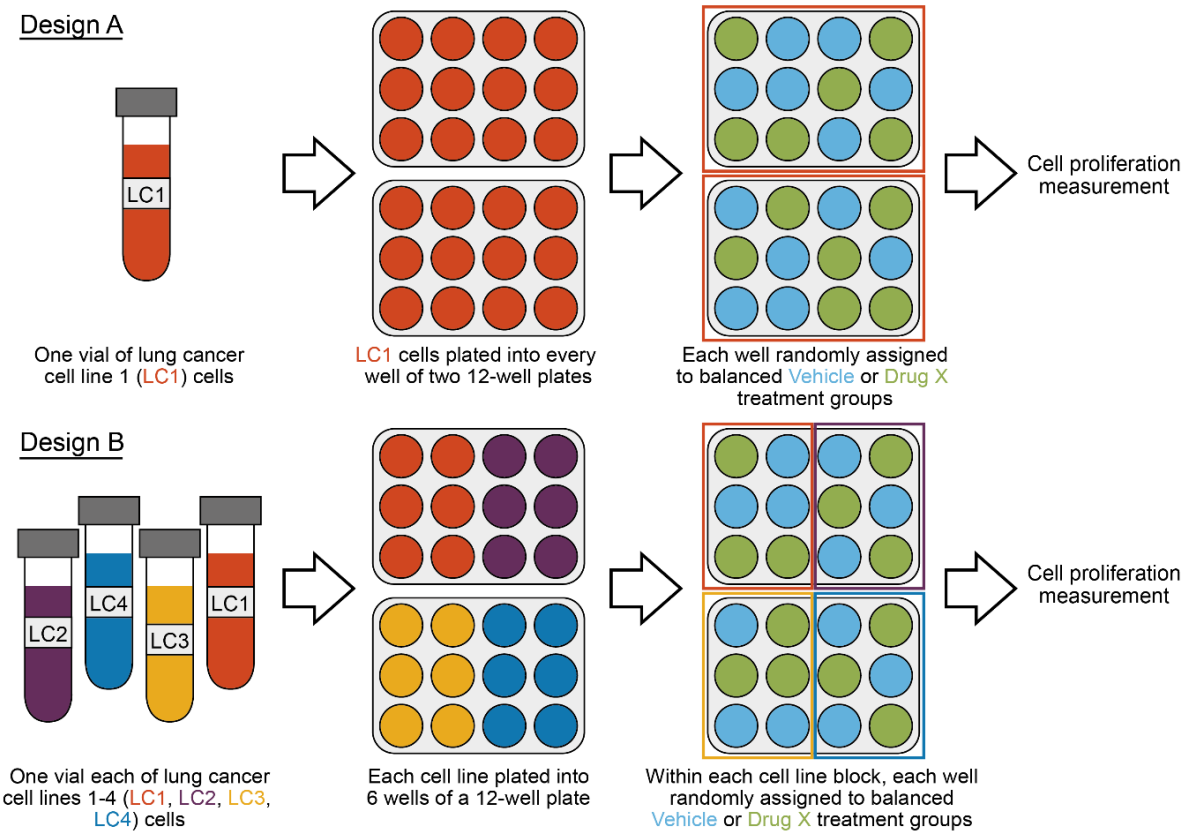

Fig 2. Generalisability in *in vitro* experiments.

This figure shows two alternative designs for a cell culture experiment, investigating the hypothesis that a drug reduces proliferation in lung cancer cell lines. Both experiments are carried out in two 12-well cell culture plates. Design A uses a single lung cancer cell line, while design B investigates the hypothesis using four independent cell lines, randomising wells to vehicle or drug treatment groups within each cell line block. In both experiments, the total sample size (*N*) is the same (24), as is the sample size of the vehicle and drug treatment groups (n = 12 in both cases) and the experimental unit (the well). As a result, the statistical power of the experiment to investigate the hypothesis is very similar in both designs (though the power may be slightly reduced in design B if treatment effects vary between cell lines).

It is useful to consider how the difference in design impacts the generalisability of the results obtained. In design A, the study's results can only apply to the single cell line used (LC1). However, the use of multiple cell lines in design B means that the results of this experiment can be generalised more broadly, making it more likely that they are applicable to *in vitro* lung cancer models in general.

In addition, the use of only a single cell line in design A makes this experiment more vulnerable to the introduction of bias, as any particularities of the single biological unit used (the vial of LC1 cells), such as bacterial contamination, will affect the entire experiment. The use of multiple biological units in design B alleviates this risk. The potential for bias to be introduced could be

further reduced by plating the four cell lines in design B across both plates in alternating columns, to account for any positional effects. However, the practicalities of implementing this arrangement while also fully randomising wells should be considered when designing the experiment (see Recommendation 2).

The sample size, genuine replication and pseudoreplication

Correctly identifying and reporting the experimental unit helps to ensure that an experiment is being genuinely replicated. Genuine replication is the process of performing the same experimental procedures on (and taking observations from) multiple independent experimental units. Because the number of experimental units is equal to the sample size, genuine replication increases the sample size of an experiment [30, 31]. In contrast, replicating experimental procedures on non-independent units, but treating these non-independent observations as if they were independent, is known as pseudoreplication or subsampling. Pseudoreplication does not increase the sample size of an experiment, even if there are more data points. Rather, it underestimates the true variability in a study, increasing the chance of both false-positive and false-negative results [32, 33]. In the example discussed previously (in which the experimental unit is a well of cells, as in figure 1A) pseudoreplication would occur if measurements were taken from individual cells in the same well and each of those measurements was treated as an independent sample. Multiple measurements taken from the same experimental unit (in this example, from individual cells in the same well) are classified as subsamples, or pseudoreplicates. Subsamples can be an important component of some experimental designs. For instance, they can be used to obtain more precise estimates of an outcome measure from a single experimental unit when measurements are noisy. However, they do not contribute to the sample size. For example, if the outcome measure in the previous example was cell size, a number of measurements could be taken on individual cells within the same well, with the mean of these results used as the single data point for that well in the analysis (alternatively, multilevel models that appropriately account for the structure of the data can be used [34]). While subsamples do not increase the sample size (and therefore do not increase the statistical power) of an experiment, they provide important insight into heterogeneity of responses, and the associated variability should always be transparently reported (see Recommendation 6).

The prevalence of pseudoreplication and inadequate reporting of the experimental unit have been noted in *in vitro* studies [35-37]. Researchers frequently use the term ‘biological replicates’ to refer to genuine replication, and ‘technical replicates’ to refer to subsampling or pseudoreplication. However, these terms are used ambiguously and inconsistently in the published literature. Repeating an experiment on a given biological entity may not correspond to genuine replication, and technical replication (however defined) may not equal pseudoreplication. The terms ‘biological replicate’ and ‘technical replicate’ do not capture the important characteristics of an experiment but can blur important distinctions, and can be used inappropriately to justify a poor experimental design [30, 31]. It is therefore recommended to avoid these terms, and use the terms experimental unit, genuine replication, and pseudoreplication, as these directly define the concepts of interest.

For each experiment, it is important to identify the experimental unit before undertaking the study, so that sample sizes can be properly determined, ideally through power analysis. Likewise, it is important to clearly identify the experimental unit in a manuscript, so that statistical analyses can be properly evaluated by a reader. This information is particularly useful when the experimental unit is distinct from the biological unit of interest or observational unit, and when there are multiple types of experimental units in a study. Manuscripts containing multiple experiments can report the experimental unit for each one using a table, placed either in the body of the manuscript or in supplementary material permanently associated with it.

Recommendation 2: Risks of bias

Report whether and how risks of bias were considered and addressed in the design of each experiment

In biological research, bias occurs when errors are introduced into the experimental process, resulting in systematic deviation between the results or conclusions of a study and the "truth" [38, 39]. For a reader to be able to evaluate the reliability of a study, it is important that any potential sources of bias present in each experiment, and any measures that have been put in place to minimise their impact, are understood and reported. The risks of bias present in a study directly impact its reliability, lower risks of bias will result in a study that is generally more reliable.

Bias can be introduced at all stages of the research process, especially during experimental design, while conducting the experiment, when measuring the results, performing data analysis, and interpretation and reporting of the results. The potential for bias to influence the results of experiments can be reduced when researchers are aware of the factors that can introduce bias to *in vitro* studies, so that steps can be taken to address those risks during the design, conduct, and reporting of an experiment.

Considering and addressing sources of bias

Researchers can inadvertently affect the outcome of an *in vitro* experiment in many ways, but a variety of techniques can be used to avoid or minimise the impact of bias. The types of bias affecting *in vitro* research can be organised into a number of categories – here these are classified as allocation bias, performance bias, detection bias and outcome reporting bias. Each of these sources is discussed in detail below. The sources of bias have been classified in this manner to align with the way that biases affecting experimental *in vivo* and human research are typically categorised. These categories may not necessarily align with the way bias in *in vitro* studies has been classified elsewhere, but their content incorporates the questions asked in other similar systems.

A number of strategies exist to mitigate the impact of these biases, including randomisation, blinding (also known as masking), and prespecifying hypotheses, statistical plans, and criteria for inclusion or exclusion of data points in analyses. These are discussed below, with regard to how each can be used to address specific sources of bias. Some techniques can be used to address more than one source of bias, these are therefore discussed across different sections. For a reader to understand the extent to which bias may have influenced the results of a study, it is essential that details of whether and how these mitigation techniques were implemented are reported.

The RIVER recommendations are not designed to be used as a tool for evaluating risks of bias in *in vitro* studies. Multiple tools designed for this purpose have been developed or proposed, including the US National Toxicology Programme's Office of Health Assessment and Translation (OHAT) approach for conducting literature assessments [40], ORD staff handbook for developing IRIS assessments [41], the ToxRTool for Toxicological data reliability assessment [42] and the SciRAP tool [43]. Ideally, tools evaluate risk of bias separately from reporting quality, consistent with current best practices [44], and are most often used as part of conducting literature-based reviews, to translate primary research findings into knowledge used to inform decision making.

Allocation Bias

Allocation bias concerns the way that experimental units are allocated to experimental groups (or treatment conditions). Allocating experimental units in a non-random manner has significant potential for introducing systematic differences into those groups, with consequent effects on the results obtained [30]. For example, in a cell culture experiment, arbitrarily assigning the wells of a culture plate to control or intervention groups according to their layout (e.g. assigning the left half of the plate to one group, right half to another) or the order they are treated (e.g. first a drug solution is added to half of the wells, then its vehicle is added to all remaining wells) can result in the introduction of differences between groups that are unrelated to the experimental intervention. These could include spatial effects from the position of samples on the plate, differences in interventions provided due to degradation of reagents over time, or different volumes of sample being added due to volume drift in pipettes. Any differences in outcome measures observed between the two groups may therefore represent differences in the groups themselves, or differences introduced by the technical conduct of the experiment.

Allocation bias can be directly addressed by using randomisation when allocating experimental units to groups. This ensures that those groups are, at the beginning of the experiment, as similar to each other as possible [30]. Haphazard or arbitrary allocation is not truly random. Using an appropriate method of randomisation ensures that each experimental unit has an equal probability of being assigned to any group, meaning that researchers' conscious and unconscious biases cannot influence allocation. Examples of suitable methods include generating random numbers (e.g. by using the Rand() function in spreadsheet software, or using online random number generators), picking numbers out of a bag, rolling a dice or flipping a

coin. Randomisation makes experimental results easier to analyse and interpret, because alternative explanations for differences between treatment groups become less plausible.

The use of inferential statistics (such as t-test or ANOVA) to analyse data is based on the assumption that experimental units are randomly allocated to treatment groups from a homogeneous background population, meaning that statistical analyses may not be valid on groups which have been allocated in a non-random manner [45].

Performance Bias

Performance bias refers to systematic differences introduced as a consequence of the way experimental groups are handled during an experiment. This can result from researchers knowing the group each experimental unit belongs to (e.g. treatment or control group), or from differences in experimental conditions between groups.

Because researchers often expect a particular outcome, knowing an experimental unit's group allocation can result in differences (intentional or unintentional) in the way that different units are treated [46]. For example, if a researcher expects an intervention (such as a drug treatment) to have a positive effect on the growth of cells in culture, this may result in the researcher handling or treating the cells in the drug treatment group with more care, or paying closer attention to the conditions in which they are cultured, to the extent that this can affect the results of the experiment.

This source of bias can be addressed by the use of blinding (also known as masking). Blinding refers to concealing information about group allocation, and other aspects of the experiment, from the researchers carrying it out; this can be done at various stages of an experiment, reducing the likelihood of unconscious biases – such as the expectations of a researcher – influencing experimental results [46]. It may be difficult to mask group allocation throughout the entire experiment, because of the experimental design or other practical factors; however, it should always be possible to do so at some stage(s). As well as the group allocation of each experimental unit, other important information to conceal includes any relevant properties of a sample, such as whether it was derived from a healthy or diseased source, along with any information that may provide clues about group allocation such as the date of sample collection (for example, if disease samples are only collected on specific dates).

Blinding can be implemented by various means. For example, in a cell culture experiment (such as that shown in figure 1A) one researcher could randomly allocate each well of a cell culture plate to control or drug treatment, then either mask or code the labels of containers with the treatment and control solutions, such that the only information on each container is the well to which it is assigned. A second researcher, responsible for carrying out the intervention, can then apply the solutions to each well without being aware of which wells are receiving a control and which are receiving a drug treatment. Thus, the second researcher's expectations cannot influence the way they handle different experimental units or groups during the experiment.

Performance bias can also result from systematic differences in experimental conditions between groups. These differences can arise from the properties of the experimental method, model, or processes being employed. For example, conditions between groups could vary systematically as a result of the order in which samples are subjected to interventions, the day, time or order in which sample outcomes are measured, the position of samples on plates, arrays, chips or gels, the source of samples or reagents (when they derive from different vendors or healthcare settings, for example) and the limitations of any equipment used.

When factors with the potential to introduce systematic differences are identified, they can be accounted for in the design of the experiment. One strategy is to randomise or alternate samples to address the specific issue identified. For example, in cell culture experiments, wells on the periphery of a plate are often less well humidified than wells towards the centre, meaning that proliferation of cells in peripheral wells may be impaired relative to others. To address this, the position of samples from different groups can be randomised (either simultaneously with random allocation to groups or after allocation has taken place, depending on the experimental design) or alternated. Another strategy is to keep these factors constant for all groups. For example, in an experiment involving DNA extraction from tissue specimens, processing samples in more than one batch may introduce differences between batches, which could be misinterpreted as differences between groups and introduce bias (particularly if some experimental groups are overrepresented in some batches) [30]. This can be addressed by processing all the samples in a single batch, such that all groups are treated under the same conditions. Alternatively, if logistical considerations mean that DNA extraction must be carried out in multiple batches, samples from each experimental group can be split equally across each batch, so that each group is equally exposed to any differences caused by sample processing. These strategies can also be used together: in this DNA extraction example, randomising or alternating the order in which samples from different groups are processed (within each batch) also addresses any differences that could be introduced by sample processing order (for example, resulting from reagent degradation over time).

Detection bias

Detection bias refers to systematic differences between experimental groups that are introduced as a consequence of how outcomes are assessed. This can again result from researchers' knowledge of experimental units' group allocation, when researchers' expectations can affect how outcomes are recorded and analysed [47], or from limitations in the methods or equipment used to measure or analyse outcomes.

As with performance bias, detection bias can be addressed by the use of blinding when measuring outcomes or during data processing and analysis. This is especially important for subjectively measured outcomes, such as visual assessment of apoptotic cell numbers in culture, counting cells in a proliferation assay using a vital dye (e.g. Crystal Violet) or observing the degree of pathology or differentiation in tissue samples. Studies examining the reproducibility of preclinical research (including *in vitro* studies) have found a correlation between irreproducible studies and a lack of blinding during data analysis [48].

Samples' group allocation can be masked during outcome assessment in numerous ways. For example, in an experiment assessing histological changes in explanted skin samples (such as that shown in figure 1C) one

researcher could re-code the labels on images of stained samples, such that a second researcher assessing the histological changes is unaware which samples belong to which group. The risk of detection bias can be further decreased by ensuring that outcome measures (particularly subjectively measured ones) for entire groups are not assessed in consecutive order (i.e. by not assessing all control group samples followed by all intervention group samples). Instead, the order of outcome measurements can be randomised, or samples from one group can be alternated with samples from others, such that the researcher assessing the outcome(s) is not aware, from the order of samples alone, whether consecutive samples belong to the same group.

Separately, detection bias can be mitigated by making use of methods or equipment that are sensitive and appropriate to the range of measures expected in the experiment. For this purpose, it can be useful to consider certain key questions related to the experimental setting, such as:

- How accurate is the method?
- What is the range of measurements that provides reliable data?
- What is the variability associated with the measurements conducted?
- What is the sensitivity of the method or equipment used, relative to that range?

In this context, ‘sensitivity’ refers to the smallest absolute amount of change that can be reliably detected by a measurement (i.e. the lower limit of detection). Similarly important is the upper limit of detection that provides precise and accurate results. An *in vitro* method, including the equipment used as part of it, should be developed and validated in such a way that the expected values for each outcome measure fall between these defined lower and upper detection limits [19]. The use of inappropriate equipment, unsuitable for detecting values in the expected range of measurements, therefore represents a potential source of bias, as it may produce results systematically higher or lower than would be expected [18, 49].

For example, when quantifying western blots, an important factor to consider is the linear range of detection for the protein of interest, as well as the loading control used for normalisation. The linear range of quantification is the region in which there is a linear relationship between the amount of target protein on the membrane and the detected result. As the amount of protein on the membrane increases, so should the signal intensity. If the signal is outside this range, it either indicates low sensitivity, where signals may not reflect the true sample concentration, or an excessively strong signal, which may cause membrane, film or detector saturation. This can affect the detected difference between groups if all samples are not within this range. For instance, if one treatment group consistently results in saturated signal, differences between groups will appear smaller than they are in reality. The goal, therefore, is to avoid saturation and low sensitivity by designing the experiment such that signal remains in the linear range, as this will provide the most reliable and reproducible results, and avoid the introduction of any systematic differences [20].

When assessing how to address the risk of detection bias, it is useful to evaluate whether automated techniques, which would reduce the risk of bias associated with, for example, subjective assessment of experimental outcomes, can be used. For example, in an experiment assessing changes in cell size, use of

automated imaging software to calculate cells' area could avoid the potential for bias to be introduced by manual measurement.

Outcome reporting bias

Outcome reporting bias refers to selective or distorted reporting of experimental results, and/or biased interpretation of available information. This type of bias can result from a number of practices, including selective reporting, 'p-hacking' and hypothesising after experimental results are known ('HARKing').

Selective reporting refers to the inclusion of only some experimental outcomes, analyses or groups (for example, reporting only 'positive' results). This can introduce bias to a study, as the results presented are not representative of the actual data generated by experiments. This can result in over- or underestimation of the true effect of experimental interventions, and systematic deviation between the results reported and the true results of experiments. This practice can be avoided by reporting all the results generated from every experimental group, which is discussed in more detail in Recommendation 5.

P-hacking (also known as data dredging) involves analysing a dataset in multiple ways to identify statistically significant results. This includes running different statistical tests; analysing data repeatedly during collection; combining, splitting or excluding experimental groups; or selecting the outcome measures that produced significant results [50]. HARKing refers to the process of generating hypotheses *post hoc* (i.e. after experiments take place), but reporting them as *a priori* hypotheses that were confirmed by the results of the experiment [51]. While generating hypotheses is the genuine purpose of some exploratory experiments (for example, single-cell sequencing or other observational studies), HARKing only refers to experiments inaccurately presented as having been performed to test these *post hoc* hypotheses.

Both p-hacking and HARKing give a misleading impression of how regularly scientific hypotheses are confirmed by the results of experiments investigating them [52]. These practices introduce bias in entire research fields by increasing the probability of false positive findings being published, and reducing the reproducibility of experimental results. This can be addressed by predefining the analysis plan which sets out how an experiment will be designed, and the analytic workflow that will be employed. This ensures that experiments are designed to address the specific research question of interest and reduces the temptation to select the analysis pathways or change the hypothesis based on the results observed. An analysis plan will typically include:

1. The aim of the experiment or the hypothesis being tested.
2. The design of the experiment.
3. Primary and secondary outcome measures.
4. The treatments or experimental conditions.

5. Any data preprocessing steps to be carried out, such as transformations, normalisations, and quality control checks.

6. Criteria for inclusion or exclusion of data points in analyses (e.g. technical acceptance criteria).

7. The analysis, test, or model to be used, and its applicability to the experimental design and type of data collected.

Prespecifying criteria for including or excluding data points, experimental units (e.g. samples) or groups from analyses reduces the capacity for researchers to influence results by selecting whether or not to include particular measurements in those analyses. For example, these criteria may define which measurements can be considered to be 'outliers', or may define the criteria by which an experiment can be judged to be a technical failure. The importance of setting and reporting criteria for including or excluding data points, experimental units or groups from analysis is discussed in more detail in Recommendation 5.

An experimental plan can be preregistered (submitted to a journal or registry) before experiments are carried out [53]. One version of this process is known as a registered report, in which a study's proposed methods and analyses are submitted to, and peer reviewed by, a journal in advance of the experiment taking place. Publication of the study is then dependent only on the rigour of the methodology and the development of the scientific question, rather than its results [54].

Identifying opportunities to address bias

Figure 3 shows examples of questions that are useful to consider when assessing whether and how bias can impact a study. Considering these during the design of an experiment can help to identify the types of bias that may be introduced, and the techniques that could be used to address them.

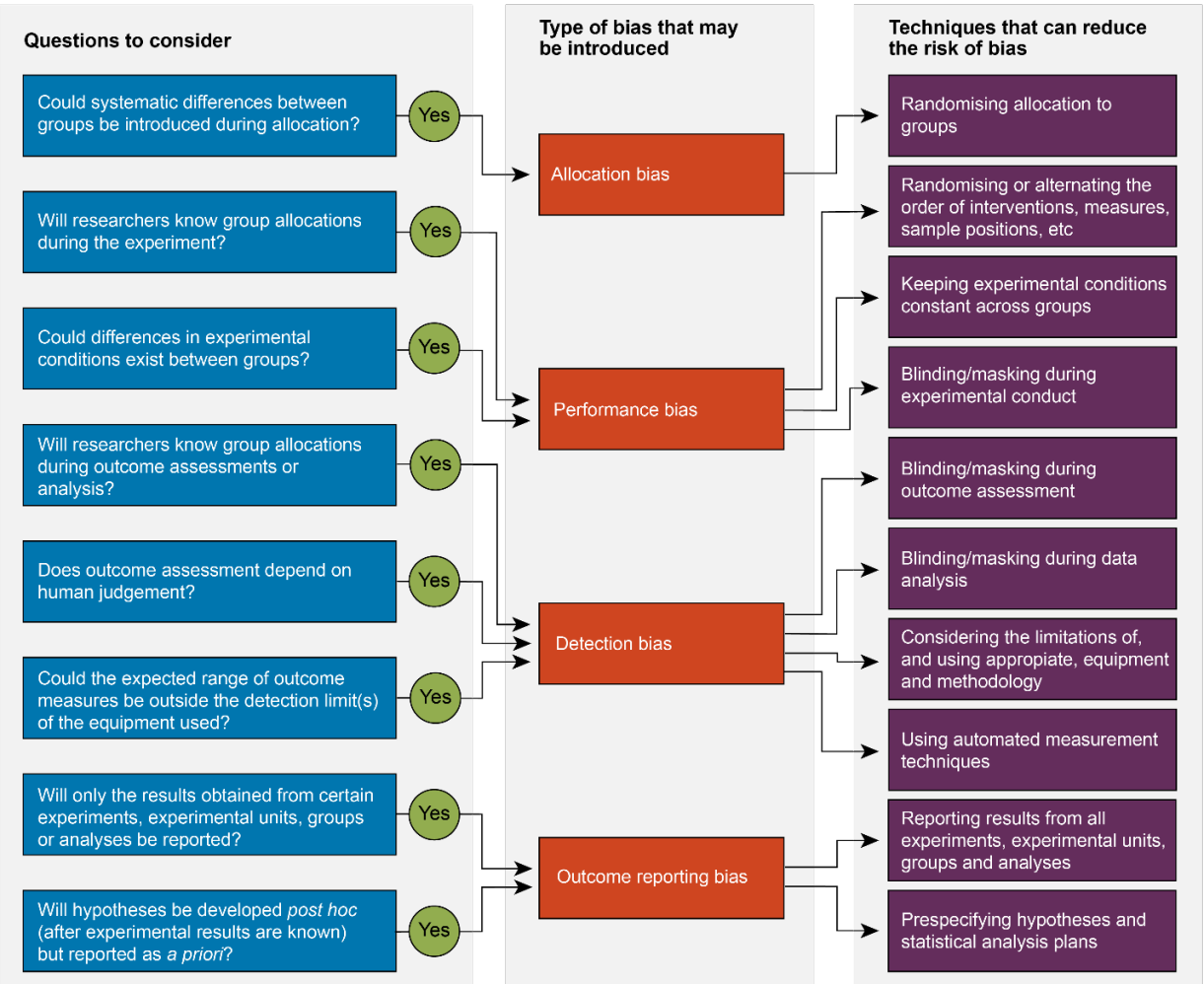

Fig 3. Diagram illustrating questions that may be useful to consider (blue) when identifying and addressing risks of bias in *in vitro* research.

The diagram shows the type(s) of bias that may be introduced in regard to each question (orange), and the techniques that can be used to address each type of bias (purple).

It is helpful to consider the balance between the protection from bias offered by each technique, and the practicalities of implementing the measure. While implementing measures to reduce the impact of bias represents best practice and it should always be possible to implement some of the above-mentioned techniques, there can sometimes be justifications for not using these techniques in an *in vitro* study. For example, completely randomising 96 samples across a microtitre plate may not be feasible, as it can potentially introduce a high number of pipetting errors and unsustainably increase the time required to complete an experiment. In addition, strategies such as blinding (masking) may not be necessary in experiments for which internal controls exist. For example, in an experiment measuring gene expression by quantitative reverse transcriptase PCR (qRT-PCR) using absolute quantification, concealing samples' group allocation may not be necessary, as the reference standards included in the experiment (provided these are

included on the same plate as the experimental samples) allow for objective, absolute measurement of target RNA concentration. It may also not always be appropriate to predefine a specific hypothesis, particularly for exploratory research such as sequencing studies to identify genomic, proteomic or metabolomic responses to chemical exposures.

For a reader to understand the degree to which bias may have influenced a study's results, it is important to report the sources of bias that have been considered, and the measures that were used to reduce the risk that they impact on the findings. If there were justification for not implementing such measures, or if no measures addressing the identified sources of bias exist, including this information in a manuscript will help readers assess how trustworthy the experimental findings are.

Recommendation 3: Experimental model

Provide details of the model used in each experiment, including (if applicable) the identity, source and life stage of any cells used, information on microenvironmental (i.e. physical and biochemical) conditions, and quality control metrics.

In an *in vitro* experiment, the experimental model is the system that is used to investigate the research question. In an *in vitro* context, an experimental model may be cell-free (for example involving the direct interaction between proteins and small molecules), or involve the use of living cells, microorganisms, or tissues derived from human, animal, plant or other living sources. Models involving living cells may place them in adherent or suspension culture, or in more complex three-dimensional systems, such as organs-on-a-chip, organoids, microfluidic systems or within scaffolds, among many other approaches. It is important to report details of the model used in each experiment, so that a reader can fully understand the approach taken to investigate a study's hypothesis and assess its suitability [18]. Reporting all relevant details of a model also allows readers to reproduce that model to investigate similar hypotheses, or adapt it for the study of other research questions.

When reporting the use of complex models, a diagram or schematic of the model system may represent the most useful means of communicating its design and structure. Important details of each model to report include:

Experimental setup:

- The identity of any cells or cell lines used, including their source, and whether and how they have been characterised and/or authenticated (e.g. by karyotyping or genetic testing) [55, 56]. It is also useful to describe why specific cells were chosen (for example, whether they express a target protein of interest).

- For human- or animal-derived samples, induced pluripotent stem cells (iPSCs), organoids or similar models, as well as established cell lines, details of the species, sex, age, ethnicity and other relevant characteristics of the source – human or animal. Where human samples are used, consideration should be given to ensuring these details cannot be used to identify specific individuals. Details of relevant ethical approvals should also be provided.
- Details of how cells or samples were managed in advance of the experiment, including extraction processes, the temperature [57], CO_2 concentration and humidity at which they were cultured, the composition of the medium in which they were maintained (including whether antibiotics or antimycotics were used), the duration of their maintenance, mycoplasma status [58], number of population doublings [59], and their confluency at the start of the experiment [60, 61].

Experimental conduct:

- Details of the physical and chemical environment in which the experiment took place. This includes details of the physical environment's design (for example, whether the experiment took place in suspension or adherent cell culture plasticware, microfluidic chips, hydrogels, organoids, using structural scaffolds, or other physical environment), and the temperature, CO_2 concentration, humidity, and composition of the medium or substrate in which the experiment was performed.
- If cells were used, details of whether and how any measurement or assessment of cell viability and cell number, or analysis of cells' phenotypic stability in the model took place.
- Details of any quality control metrics or assessment which made up part of the experimental model. This applies to not only the biological material used (for example, cells) but also to reagents or materials used to generate the experiment model (for example, devices, substrates or reagents).
- Details of any prior validation of the model or method. This could include characterisation of the model performed by the researchers themselves, previous publications demonstrating the applicability of the model to the research question, or information on acceptance of the model as part of published international test guidelines.

For studies describing novel or adapted experimental models, it is important to include information on how the model was developed, characterised and/or benchmarked. This is especially true for studies describing complex *in vitro* models, such as three-dimensional cell cultures, organs-on-a-chip or organoids. Depositing detailed protocols for how such models are generated in openly accessible repositories (see Recommendation 4) is also strongly encouraged.

Recommendation 4: Experimental procedures

For each experiment, describe the experimental procedures in detail, including what was done, when and how often, and using which equipment and reagents.

When describing experimental procedures, important details to report include what was done, when and how often, and which equipment and reagents were used. Including this information in a manuscript helps to ensure that readers can understand exactly how each experiment was performed and how to replicate it, without the need to consult external sources for essential information. Depositing detailed, transparent, and versioned experimental protocols in publicly available protocol repositories (examples include [62-64]) is strongly encouraged, as this enhances the reproducibility of experimental procedures by providing full methodological transparency, and allows protocols to be updated as they are further developed over time. When describing complex procedures, an accompanying diagram, flowchart, or similar graphical representation may represent a clearer means of displaying experimental procedures than narrative text. References to published protocols or papers describing similar procedures are a useful way to provide context to reported methods. However, as discussed previously, important information is frequently missing from published references, and this is not an acceptable substitute for providing enough information in the manuscript itself or in a publicly-available protocol.

Of particular interest are the test articles used in an experiment. Test articles are any reagents used as interventions or treatments during an experiment, as well as substances used as positive, negative, or vehicle controls. Important details to report for each test article include the unique identifier, source, purity, dosage, concentration, and duration of any use. When reporting the use of test articles, as well as other reagents and resources (including antibodies and cells), Research Resource Identifiers (RRIDs) – unique, stable digital identifiers that allow unambiguous identification of reagents or tools used in a study [65, 66] – are especially useful to include, as these aid other researchers in replicating published methods. If chemical substances are used, unique chemical identifiers such as InChI [67] provide clear detail to readers on the precise substances used. For experiments using antibodies, it is important to report whether and how they were validated for use in the assay(s) performed. For example, when using newly generated antibodies, presenting appropriate controls demonstrates that the antibody specifically detects the antigen of interest. It is useful to report the lot number(s) of any biological reagents used, including serum and bovine serum albumin (BSA) among others, as well as antibodies. If more than one lot is used in the same experiment, it is important to consider and test whether this could introduce bias to the results (i.e. whether different lots have different experimental effects in practice; see Recommendation 2).

When describing experiments using specific equipment, it is useful to include information on calibration, use of reference standards and relevant hardware or software versions. For experiments relying on robotics and/or imaging technology, important details to report include details of the instrumentation and software packages used, any custom scripts used during analyses, and any post-acquisition data processing steps. Any scripts and code used should, along with source data, be archived for future reference in a FAIR-

compliant repository [68] (see Recommendation 6) capable of minting DOIs, with an appropriate link reported within the methods section, and/or fully reproduced in supplementary information associated with the manuscript.

Recommendation 5: Experimental groups and exclusions

Report all data obtained from all experimental groups (including controls) and justify any exclusions

Data exclusion refers to the removal or disqualification of data points, experimental units or groups from an analysis. For example, this could include the removal of ‘outliers’ from datasets, however these are defined. It is important to report any data exclusions that take place, and justify them. Providing the reasoning behind the exclusion of data points enables a reader to understand how data were manipulated during an experiment, evaluate how robust the results presented are, and recognise when data have been managed responsibly. Many legitimate reasons for excluding data points exist, however excluding data arbitrarily, or according to researchers’ subjective opinions on their validity, has the potential to introduce bias to experimental results [69]. That is, when the inclusion of data in analyses is solely at researchers’ discretion, decisions on the inclusion or exclusion of particular data points can be influenced by expectations or preconceptions. This is particularly an issue in experiments where researchers are aware of the treatment each sample has received (see Recommendation 2). Excluding data is sometimes undertaken to ensure that the most compelling data are presented, rather than presenting data that truly reflects the spectrum of experimental outcomes, and it is referred to as ‘cherry-picking’. Many statistical analysis methods are extremely sensitive to outliers and missing data [70], meaning that these exclusions also have the potential to directly affect the outcomes of analyses.

To minimise the potential for bias introduced by exclusions, it is important that criteria for the inclusion or exclusion of data points or experimental units are defined before an experiment starts. These criteria should be as objective as possible, include thresholds and rules for when data should be included or excluded, and be applied equally to all experimental groups [69]. A variety of grounds for excluding data exist, which can be set out in these pre-specified criteria. These include technical failures, in which one or more aspects of a model or experiment fail to meet pre-specified criteria for validity (for example, samples may fail to meet specified quality control standards, such as by being of insufficient volume, contaminated, or of poor quality by another measure); equipment failures, such as errors that produce measurements outside plausible ranges (or no measurements at all); or data handling failures, such as obvious errors in data entry that produce clearly inaccurate results. Exclusion criteria can be included in a statistical analysis plan defined in advance of an experiment (see Recommendation 2) and should be reported in manuscripts.

In addition to exclusion of data points, the selective (or incomplete) reporting of experimental results or analyses, in order to support a particular hypothesis (i.e. only reporting 'positive' results), has a considerable impact on the reliability and reproducibility of those results. Preference for positive results has often been documented; studies demonstrate that statistically significant results are more likely to be submitted, published and cited [71, 72]. Some consider that presenting every experimental replicate and result generated during a project conflicts with the narrative structure commonly employed in manuscripts [69, 73]. However, in an *in vitro* context, selective reporting can include omitting some outcome measures, experimental groups, replications or entire experiments from reports. If enough experiments are carried out, it is inevitable that some will obtain statistically significant results by chance alone [74]. Selectively reporting only those experiments (or parts of experiments) with significant results that support a particular hypothesis misrepresents the true rate of confirmed hypotheses, and can lead to over- or underestimation of the effect of interventions [1, 75]. For example, if a cell culture experiment investigating the effect of a drug treatment on cancer cell growth is repeated three times, two of these experiments may show non statistically significant results, and one may show a significant decrease in growth rate. Reporting only the experiment resulting in a statistically significant change can result in overestimation of the effect of the drug by readers. Future research building on these reported findings, or meta-analyses synthesising the results of published studies, will therefore be based on incomplete or misleading results.

For a reader to assess the reliability of the results reported, it is crucial that all the data obtained from each experimental group are reported, including results that are not statistically significant (or results that do not support the study's hypothesis). This includes control groups (negative or positive) included in the experimental design, and descriptions of whether positive controls had their intended effect.

Recommendation 6: Data availability and presentation

Share and present data transparently

Data Sharing

Whether and how experimental data are made available significantly impacts the utility of a study. Scientific data can be defined as: recorded, digital information created from research activities, such as experiments, analyses, measurements or observations, or resulting from simulations or models. They do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens [76, 77].

Many funders, journals and other organisations have published guidance and standards for when and how to share source data, see the following references for more information: [78-82]. Sharing adequately annotated data transparently and in compliance with applicable standards promotes reuse and verification of those data,

and reduces any ambiguity in interpretation arising from the way in which results are presented [83, 84]. It allows readers to fully understand what occurred in an analysis and replicate it, allowing results to be independently tested and verified [85]. Data sharing also allows experimental data to be repurposed, and new datasets to be created by combining data from multiple studies for use in secondary analyses, increasing the impact of the study from which the data originated [86]. Sharing source data is associated with an increased citation rate [87, 88] and improvements in researcher reputations [89], whilst providing a detailed plan for sharing data increases the perception of a grant proposal's impact [90].

Following the FAIR guiding principles [68] when sharing source data ensures that these data are Findable (e.g. among other things, they are assigned unique, persistent identifiers), Accessible (e.g. they are retrievable using their unique identifiers, and do not use outdated file types), Interoperable (e.g. they use a formal, accessible, shared, and broadly applicable language for knowledge representation) and Re-usable (e.g. their attributes are richly, accurately and relevantly described). Data can be made available via a wide range of public and journal-specific repositories, see the following references for more information: [91-96]. When data are shared via a repository, it is helpful to share links to these repositories within the manuscript, whilst statements such as ‘data are available on request’ are unhelpful to a reader and are discouraged [97].

Data presentation

The same set of data presented in different ways can lead to different interpretations. When data are presented using graphs, the type of graph, its layout, the data intervals displayed in the axes, specific data comparisons, and the presence or absence of individual data points, error bars and information on statistical significance, can all strongly affect how a dataset is interpreted [98]. The way data are presented should enable the reader to have confidence in the statistical analyses performed. For example, research has shown that most papers present continuous biological data using bar or line graphs, which present significant issues for data interpretation because many different underlying data distributions can appear identical on these graph types [99, 100]. In contrast, papers rarely present data using visualisations that allow readers to critically evaluate the distribution of the underlying data, such as histograms, scatter plots or box plots. These more complete forms of data visualisation significantly aid readers’ interpretation of the data generated in an experiment. In experiments with small sample sizes, univariate scatterplots represent the best choice for showing the distribution of data [99]. Irrespective of the graph type used, it is beneficial to present the individual data points for each experiment on graphs [98], in addition to presenting summary data (such as mean and standard deviation), and to include the sample size, the statistical method and a measurement of confidence level. In considering this recommendation, note that it may not be practical to present individual data points when large datasets are plotted. In this case, the use of visualisations that display the full degree of statistical variation in grouped data, such as box-and-whisker or violin plots, are useful ways to transparently present data. Information on best practices in data presentation can be found elsewhere, such as in the following references: [99-102].

When presenting representative images, such as those derived from imaging or immunoblotting experiments, it is important to present images in as much context as possible, including supplying unprocessed and uncropped original images of gels and western blots [103], and providing details of the number of times a technique was carried out to generate the representative images presented. Image manipulations or 'enhancements' should also be avoided. These include, for example, adjustments of brightness, contrast, or colour balance that obscure, eliminate, or misrepresent any information, including non-specific signal or background.

Conclusion

For *in vitro* experiments to add to the scientific knowledge base, it is essential that they are reported transparently and accurately. By focusing on a small number of the highest priorities, the RIVER recommendations provide a practical solution to improve the reporting of *in vitro* experiments. The concise list will facilitate adoption by journals and other key stakeholders, and allow rapid progress.

The RIVER recommendations are not intended to supersede other *in vitro* research standards, such as regulatory requirements or individual journals' guidance – they are designed to work alongside and complement them. While the recommendations provide guidance to help communicate *in vitro* research effectively and transparently, they also introduce concepts that will promote more robust study designs in future research.

Transparent and accurate reporting benefits the whole scientific community, and improving reporting is a collective endeavour. We solicit and welcome feedback on the intent, design, and content of the recommendations. RIVER will also be subject to extensive user testing by researchers preparing *in vitro* manuscripts to ensure that the recommendations and their explanations are well understood and can be applied in practice.

Glossary

Experimental unit

The biological entity that is randomly and independently assigned to different experimental groups (treatment conditions) during an experiment. For statistical purposes, the number of experimental units in each group is equal to the sample size of that group, *n*.

Observational unit

The biological entity on which measurements are taken during the experiment, which may be distinct from the experimental unit.

Biological unit of interest

The biological entity that a researcher wants to make an inference or confirm a hypothesis about (e.g. an animal, cell or organelle).

Genuine replication

Replication of experimental procedures on multiple independent experimental units, which therefore increases the sample size.

Pseudoreplication/subsampling

Replication of experimental procedures on non-independent units, which does not increase the sample size even if it increases the number of data points.

Experimental model

The system used to investigate a research question. For *in vitro* experiments, this may be cell-free or involve the use of living cells, microorganisms, or tissues derived from human, animal, plant or other living sources.

Allocation bias

Bias arising from systematic differences in the way experimental units are assigned to experimental groups (e.g. to control, treatment or intervention groups).

Performance bias

Bias arising from systematic differences occurring between groups because of the way experimental units are handled during the course of an experiment.

Detection bias

Bias resulting from the way experimental outcomes are assessed, measured or analysed.

Outcome reporting bias

Selective or distorted reporting of experimental results, and/or biased interpretation of available information.

Data

Recorded, digital information created from research activities, such as experiments, analyses, measurements or observations, or resulting from simulations or models.

Hypothesising after results are known ('HARKing')

Generating hypotheses once the results of an experiment have been obtained (*post hoc*) but presenting these as *a priori* hypotheses that have been confirmed by those results.

Data dredging ('p-hacking')

Repeatedly interrogating a dataset until a statistically significant result is found, and reporting this result as the intended outcome of the experiment.

Supporting information

S1_Development: **Development of the RIVER recommendations**

Methodology used to develop the recommendations

S1_User_Testing: **RIVER Recommendations user testing methods**

Methodology to be used to road test the RIVER recommendations and the explanations for each item (as published in preprint).

Acknowledgements

RIVER working group members

Name	Roles	Affiliation
Matthew A. Brooke (Coordinator)	Investigation, Methodology, Project Administration, Resources, Visualisation, Writing – Original Draft Preparation, Writing – Review & Editing	NC3Rs, UK
Ian Ragan (Chair)	Conceptualisation, Investigation, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing	Independent consultant, UK
Glenn Begley	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Biotechnology consultant, Australia
Jessica Creery	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Office of Science Policy, NIH, USA
Jason Ekert	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	UCB Pharma, USA
Christoph H. Emmerich	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	PAASP GmbH, Germany
Maria Hodges	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	BMC, part of Springer Nature, UK
Nicole Kleinstreuer	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM), USA
Madeline A. Lancaster	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	MRC Laboratory of Molecular Biology, UK

Stanley E. Lazic	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Prioris.ai, Canada
Nathalie Percie du Sert	Conceptualisation, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing	NC3Rs, UK
Jenny Sandström	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Swiss 3Rs Competence Centre, Switzerland
Jonathan Saxe	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Cell Press, USA
Hazel Screen	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Queen Mary, University of London, UK
Emily S. Sena	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	University of Edinburgh, UK
Sowmya Swaminathan	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Springer Nature, USA
Kristina Thayer	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Environmental Protection Agency (EPA), USA
Xiaowei Zhang	Investigation, Writing – Original Draft Preparation, Writing – Review & Editing	Nanjing University, China

We would like to thank colleagues and collaborators for their advice and useful discussions around the concept and content of the manuscript.

References

1. Hirsch C and Schildknecht S (2019). In Vitro Research Reproducibility: Keeping Up High Standards. *Front Pharmacol* 10:1484 doi: 10.3389/fphar.2019.01484

2. Gosselin RD (2021). Insufficient transparency of statistical reporting in preclinical research: a scoping review. *Sci Rep* 11(1):3335 doi: 10.1038/s41598-021-83006-5

3. Poon MTC *et al.* (2021). Temozolomide sensitivity of malignant glioma cell lines - a systematic review assessing consistencies between in vitro studies. *BMC Cancer* 21(1):1240 doi: 10.1186/s12885-021-08972-5

4. Sander T *et al.* (2022). Meta-analysis on reporting practices as a source of heterogeneity in in vitro cancer research. *BMJ Open Sci* 6(1):e100272 doi: 10.1136/bmjos-2021-100272

5. The NPQIP Collaborative Group (2019). Did a change in Nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Sci* 3(1):e000035 doi: 10.1136/bmjos-2017-000035

6. Goodman SN *et al.* (2016). What does research reproducibility mean? *Sci Transl Med* 8(341):341ps12 doi: 10.1126/scitranslmed.aaf5027

7. Prinz F *et al.* (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 10(9):712 doi: 10.1038/nrd3439-c1

8. Begley CG and Ioannidis JP (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 116(1):116-26 doi: 10.1161/CIRCRESAHA.114.303819

9. Baker M (2016). 1,500 scientists lift the lid on reproducibility. *Nature* 533(7604):452-4 doi: 10.1038/533452a

10. Errington TM *et al.* (2014). An open investigation of the reproducibility of cancer biology research. *eLife* 3:e04333 doi: 10.7554/eLife.04333

11. Errington TM *et al.* (2021). Challenges for assessing replicability in preclinical cancer biology. *Elife* 10 doi: 10.7554/eLife.67995

12. Errington TM *et al.* (2021). Investigating the replicability of preclinical cancer biology. *Elife* 10 doi: 10.7554/eLife.71601

13. Percie du Sert N *et al.* (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biol* 18(7):e3000410 doi: 10.1371/journal.pbio.3000410

14. Nosek BA *et al.* (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science* 348(6242):1422-5 doi: 10.1126/science.aab2374

15. Hartung T *et al.* (2019). Toward Good In Vitro Reporting Standards. *ALTEX* 36(1):3-17 doi: 10.14573/altex.1812191

16. Macleod M *et al.* (2021). The MDAR (Materials Design Analysis Reporting) Framework for transparent reporting in the life sciences. *Proc Natl Acad Sci U S A* 118(17) doi: 10.1073/pnas.2103238118

17. Coecke S *et al.* (2005). Guidance on good cell culture practice. a report of the second ECVAM task force on good cell culture practice. *Altern Lab Anim* 33(3):261-87 doi: 10.1177/026119290503300313

18. Pamies D *et al.* (2022). Guidance document on Good Cell and Tissue Culture Practice 2.0 (GCCP 2.0). *ALTEX* 39:30-70 doi: 10.14573/altex.2111011

19. Petersen EJ *et al.* (2022). Technical framework for enabling high-quality measurements in new approach methodologies (NAMs). *Altex* doi: 10.14573/altex.2205081

20. OECD (2018). *Guidance Document on Good In Vitro Method Practices (GIVIMP)*.

21. International Society for Stem Cell Research (2023). *The Standards Initiative*. https://www.isscr.org/standards Accessed on: 14/02/2023.

22. National Institutes of Health (2014). *Principles and Guidelines for Reporting Preclinical Research*. https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research Accessed on: 05/10/2022.

23. National Institutes of Health (2023). Enhancing Reproducibility through Rigor and Transparency. doi, https://grants.nih.gov/policy/reproducibility/index.htm

24. Lauer M (2023). Take Advantage of Our Many Resources for Enhancing the Rigor of Animal Research. doi, https://nexus.od.nih.gov/all/2023/02/10/take-advantage-of-our-many-resources-for-enhancing-the-rigor-of-animal-research/

25. Medical Research Council (2021). *Methodology and experimental design in applications: Guidance for reviewers and applicants*. https://www.ukri.org/publications/methodology-and-experimental-design-guidance/ Accessed on: 04/10/2022.

26. Vaux DL *et al.* (2012). Replicates and repeats--what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep* 13(4):291-6 doi: 10.1038/embor.2012.36

27. Festing MF (2003). Principles: the need for better experimental design. *Trends Pharmacol Sci* 24(7):341-5 doi: 10.1016/S0165-6147(03)00159-7

28. Casella G (2008). *Statistical Design*. Springer.

29. Mead RG, S.G.; Mead, A (2012). *Statistical Principles for the Design of Experiments: Applications to Real Experiments*. Cambridge University Press.

30. Lazic SE (2016). *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility* Cambridge University Press.

31. Lazic SE *et al.* (2018). What exactly is 'N' in cell culture and animal experiments? *PLoS Biol* 16(4):e2005282 doi: 10.1371/journal.pbio.2005282

32. Lazic SE (2022). Genuine replication and pseudoreplication. *Nature Reviews Methods Primers* 2(1):23 doi: 10.1038/s43586-022-00114-w

33. Lazic SE (2010). The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* 11:5 doi: 10.1186/1471-2202-11-5

34. Aarts E *et al.* (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience* 17(4):491-496 doi: 10.1038/nn.3648

35. Sikkel MB *et al.* (2017). Hierarchical statistical techniques are necessary to draw reliable conclusions from analysis of isolated cardiomyocyte studies. *Cardiovasc Res* 113(14):1743-1752 doi: 10.1093/cvr/cvx151

36. Emmerich C. *Accurate design of in vitro experiments – why does it matter?* https://paasp.net/accurate-design-of-in-vitro-experiments-why-does-it-matter/ Accessed on: 08/08/2022.

37. Zimmerman KD *et al.* (2021). A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun* 12(1):738 doi: 10.1038/s41467-021-21038-1

38. Yarborough M (2021). Moving towards less biased research. *BMJ Open Sci* 5(1):e100116 doi: 10.1136/bmjos-2020-100116

39. Pannucci CJ and Wilkins EG (2010). Identifying and avoiding bias in research. *Plast Reconstr Surg* 126(2):619-625 doi: 10.1097/PRS.0b013e3181de24bc

40. National Toxicology Program, (2019). *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration*. https://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookmarch2019_508.pdf

41. U.S. EPA Office of Research and Development, (2022). *ORD Staff Handbook for Developing IRIS Assessments*. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=356370

42. Schneider K *et al.* (2009). “ToxRTool”, a new tool to assess the reliability of toxicological data. *Toxicology Letters* 189(2):138-144 doi: https://doi.org/10.1016/j.toxlet.2009.05.013

43. Roth N *et al.* (2021). Development of the SciRAP Approach for Evaluating the Reliability and Relevance of in vitro Toxicity Data. *Frontiers in Toxicology* 3 doi: 10.3389/ftox.2021.746430

44. Higgins JPT *et al.* (2019). *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd Edition. John Wiley & Sons.

45. Altman DG and Bland JM (1999). Statistics notes. Treatment allocation in controlled trials: why randomise? *BMJ* 318(7192):1209 doi: 10.1136/bmj.318.7192.1209

46. Nuzzo R (2015). How scientists fool themselves – and how they can stop. *Nature* 526(7572):182-185 doi: 10.1038/526182a

47. Maddox J *et al.* (1988). "High-dilution" experiments a delusion. *Nature* 334(6180):287-290 doi: 10.1038/334287a0

48. Begley CG and Ellis LM (2012). Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531-3 doi: 10.1038/483531a

49. Bal-Price A *et al.* (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX* 35(3):306-352 doi: 10.14573/altex.1712081

50. Head ML *et al.* (2015). The extent and consequences of p-hacking in science. *PLoS Biol* 13(3):e1002106 doi: 10.1371/journal.pbio.1002106

51. Kerr NL (1998). HARKing: hypothesizing after the results are known. *Pers Soc Psychol Rev* 2(3):196-217 doi: 10.1207/s15327957pspr0203_4

52. Rubin M (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology* 21(4):308-320 doi: 10.1037/gpr0000128

53. Nosek BA *et al.* (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11):2600-2606 doi: doi:10.1073/pnas.1708274114

54. Centre for Open Science. *Registered Reports*. https://www.cos.io/initiatives/registered-reports Accessed on:

55. Plant AL *et al.* (2014). Improved reproducibility by assuring confidence in measurements in biomedical research. *Nature Methods* 11(9):895-898 doi: 10.1038/nmeth.3076

56. Almeida JL *et al.* (2016). Standards for Cell Line Authentication and Beyond. *PLOS Biology* 14(6):e1002476 doi: 10.1371/journal.pbio.1002476

57. Lepock JR (2005). How do cells respond to their thermal environment? *International Journal of Hyperthermia* 21(8):681-687 doi: 10.1080/02656730500307298

58. Olarerin-George AO and Hogenesch JB (2015). Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Research* 43(5):2535-2542 doi: 10.1093/nar/gkv136

59. American Type Culture Collection. *Passage number effects in cell lines*. https://www.atcc.org/resources/technical-documents/passage-number-effects-in-cell-lines Accessed on: 08/08/2022.

60. Pavel M *et al.* (2018). Contact inhibition controls cell survival and proliferation via YAP/TAZ-autophagy axis. *Nature Communications* 9(1):2961 doi: 10.1038/s41467-018-05388-x

61. Efremov YM *et al.* (2013). The effects of confluency on cell mechanical properties. *Journal of Biomechanics* 46(6):1081-1087 doi: 10.1016/j.jbiomech.2013.01.022

62. . *Bio-protocol*. https://bio-protocol.org Accessed on: 31/07/2022.

63. . *Protocol Exchange*. https://protocolexchange.researchsquare.com/ Accessed on: 31/07/2022.

64. . *Protocols.io*. https://protocols.io Accessed on: 31/07/2022.

65. Bandrowski A *et al.* (2016). The Resource Identification Initiative: A cultural shift in publishing. *Journal of Comparative Neurology* 524(1):8-22 doi: 10.1002/cne.23913

66. Bandrowski AEM, M. E. (2016). RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. *Neuron* doi: 10.1016/j.neuron.2016.04.030

67. Heller SR *et al.* (2015). InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* 7(1):23 doi: 10.1186/s13321-015-0068-4

68. Wilkinson MD *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1):160018 doi: 10.1038/sdata.2016.18

69. Neves K and Amaral OB (2020). Addressing selective reporting of experiments through predefined exclusion criteria. *eLife* 9:e56626 doi: 10.7554/eLife.56626

70. Kwak SK and Kim JH (2017). Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiol* 70(4):407-411 doi: 10.4097/kjae.2017.70.4.407

71. Dickersin K (1990). The existence of publication bias and risk factors for its occurrence. *Jama* 263(10):1385-9 doi: 10.1001/jama.1990.03440100097014

72. Song F *et al.* (2009). Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol* 9:79 doi: 10.1186/1471-2288-9-79

73. Sanes JR (2019). Tell me a story. *eLife* 8:e50527 doi: 10.7554/eLife.50527

74. Ioannidis JP (2005). Why most published research findings are false. *PLoS Med* 2(8):e124 doi: 10.1371/journal.pmed.0020124

75. Reid EK *et al.* (2015). Managing the incidence of selective reporting bias: a survey of Cochrane review groups. *Systematic Reviews* 4(1):85 doi: 10.1186/s13643-015-0070-y

76. National Institutes of Health (2020). *Final NIH Policy for Data Management and Sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html Accessed on: 08/08/2022.

77. UKRI (2018). *Guidance on best practice in the management of research data*. https://www.ukri.org/wp-content/uploads/2020/10/UKRI-020920-GuidanceBestPracticeManagementResearchData.pdf

78. UKRI. *Open Research*. https://www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/open-research/ Accessed on: 8 August.

79. National Institutes of Health. *Data Sharing Approaches*. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/data-sharing-approaches#after Accessed on: 08/08/2022.

80. Nature Portfolio. *Reporting standards and availability of data, materials, code and protocols*. https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards Accessed on: 08/08/2022.

81. PLOS ONE. *Data Availability*. https://journals.plos.org/plosone/s/data-availability Accessed on: 08/08/2022.

82. Starr J *et al.* (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Comput Sci* 1 doi: 10.7717/peerj-cs.1

83. Hewitt JA *et al.* (2017). Accelerating Biomedical Discoveries through Rigor and Transparency. *ILAR Journal* 58(1):115-128 doi: 10.1093/ilar/ilx011

84. Miyakawa T (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain* 13(1):24 doi: 10.1186/s13041-020-0552-2

85. Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. *FORCE11* doi: 10.25490/a97f-egyk

86. Stieglitz S *et al.* (2020). When are researchers willing to share their data? – Impacts of values and uncertainty on open data in academia. *PLOS ONE* 15(7):e0234172 doi: 10.1371/journal.pone.0234172

87. Colavizza G *et al.* (2020). The citation advantage of linking publications to research data. *PLoS One* 15(4):e0230416 doi: 10.1371/journal.pone.0230416

88. Piwowar HA *et al.* (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE* 2(3):e308 doi: 10.1371/journal.pone.0000308

89. Tenopir C *et al.* (2011). Data Sharing by Scientists: Practices and Perceptions. *PLOS ONE* 6(6):e21101 doi: 10.1371/journal.pone.0021101

90. Wilson SL *et al.* (2021). Sharing biological data: why, when, and how. *FEBS Letters* 595(7):847-863 doi: 10.1002/1873-3468.14067

91. National Institutes of Health. *Selecting a Data Repository*. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository Accessed on: 08/08/2022.

92. National Institutes of Health. *Repositories for Sharing Scientific Data*. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data Accessed on: 08/08/2022.

93. Nature. *Data Repository Guidance*. https://www.nature.com/sdata/policies/repositories Accessed on: 08/08/2022.

94. Registry of Research Data Repositories (Re3data). *Registry*. https://www.re3data.org/ Accessed on: 08/08/2022.

95. FAIRsharing. *Repositories*. https://fairsharing.org/search?page=1&recordType=repository Accessed on: 08/08/2022.

96. Taylor & Francis. *Understanding and using Data Repositories*. https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/repositories/ Accessed on: 8 August.

97. Hrynaszkiewicz IS, N.; Hussain, A.; Grant, R. and Goudie, S., (2020). Developing a Research Data Policy Framework for All Journals and Publishers. *Data Science Journal* 19 doi: 10.5334/dsj-2020-005

98. Nature Biomedical Engineering (2017). Show the dots in plots. *Nature Biomedical Engineering* 1(5):0079 doi: 10.1038/s41551-017-0079

99. Weissgerber TL *et al.* (2015). Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biology* 13(4):e1002128 doi: 10.1371/journal.pbio.1002128

100. Drummond G and Vowler S (2011). Show the data, don't conceal them. *British Journal of Pharmacology* 163(2):208-210 doi: 10.1111/j.1476-5381.2011.01251.x

101. Kelleher C and Wagener T (2011). Ten guidelines for effective data visualization in scientific publications. *Environmental Modelling & Software* 26(6):822-827 doi: 10.1016/j.envsoft.2010.12.006

102. Shen H *et al.* (2019). Information visualisation methods and techniques: State-of-the-art and future directions. *Journal of Industrial Information Integration* 16:100102 doi: 10.1016/j.jii.2019.07.003

103. Nature Portfolio. *Image integrity and standards*. https://www.nature.com/nature-portfolio/editorial-policies/image-integrity Accessed on: 21/11/2022.

S1_Development: Development of the RIVER recommendations

1. Development process

The RIVER working group was convened by the NC3Rs, who acted as the coordinators of the group as well as active participants. At the outset of the project, a set of guiding principles for development of the recommendations was established by the working group. These stated that:

- The recommendations should contain a small number of the highest priorities to address when reporting an *in vitro* experiment.
- The recommendations should be applicable to all *in vitro* study types. For these purposes, *in vitro* is defined as “any biological experiment not carried out using living animals”.
- The recommendations should focus on ensuring that the reliability of results presented in a manuscript can be effectively assessed. For these purposes, the working group agreed that a reliable manuscript is one in which “the information provided in the manuscript justifies the conclusions made about the data”.
- The recommendations should be accompanied by explanations that are designed to ensure researchers can understand the importance of reporting particular information.

An initial draft of the recommendations, informed by consultation with *in vitro* researchers, was developed by the NC3Rs. Using this as a starting point, the recommendations were developed and refined iteratively by the working group, via a series of discussions, virtual meetings, and collaborative online working. This process identified six potential recommendations for improving the reporting of *in vitro* experiments.

2. Prioritisation exercise – methods

In April 2022 a prioritisation exercise was undertaken, to assess how important each member of the working group (n = 15) considered the six proposed recommendations to be. The purpose of the exercise was to define which recommendations would be taken forward for further development. In this exercise, working group members considered each recommendation according to the following statement:

> “How important is this piece of information for assessing the reliability of the results presented in a paper describing *in vitro* experiments?”

Working group members scored each item from one (lowest importance) to nine (highest importance), giving no two items the same score. These scores corresponded to the following descriptions:

- 1-3 – not important.
- 4-6 – important, but not critical.
- 7-9 – critical.

3. Prioritisation exercise – results

Table 1 shows the demographics of the working group members who participated in the prioritisation exercise. NC3Rs staff and the working group chair did not take part.

Primary country of work	
USA	6
UK	4
Australia	1
Canada	1
China	1
Germany	1
Switzerland	1

Sector of work	
Academia	4
Publishing	3
Funder	2
Biotechnology consultancy	1
Government	1
Pharmaceutical Industry	1
Regulator	1
Scientific consultancy	1
Statistical consultancy	1

Professional role	
Chief Scientific Officer	1
Consultant	1
Deputy head of research integrity and author experience	1
Director	1
Division director	1
Executive director	1
Executive editor	1
Health science policy analyst	1
Head of collaborations	1
Head of complex in vitro models	1
Professor of biomedical engineering	1
Professor of ecotoxicology	1
Professor of meta-science and translational medicine	1
Research group leader	1
Statistician	1

Table 1: Demographics of the RIVER working group members who participated in the prioritisation exercise.

Table 2 shows the median scores given by the participants to each proposed recommendation. The distribution of scores awarded is shown in Figure 1.

	Recommendation	**Median score and range**
1.	Define the experimental and biological units used in each experiment	6 (4-9)
2.	Report whether and how risks of bias were considered and addressed in the design of the experiment.	5 (4-9)
3.	Provide details of the model being used in each experiment, including (if applicable) the identity, source and life stage of any cells used, information on microenvironmental (i.e. physical and biochemical) conditions, and quality control metrics.	8 (4-9)
4.	For each experiment, describe the experimental procedures in detail, including what was done, when and how often, where, and using which equipment and reagents.	7 (4-9)
5.	Report all data obtained from all experimental groups (including controls) and justify any exclusions.	5 (4-9)
6.	Share and present data transparently.	7 (4-9)

Table 2: Median score and range of awarded scores for each proposed RIVER recommendation.

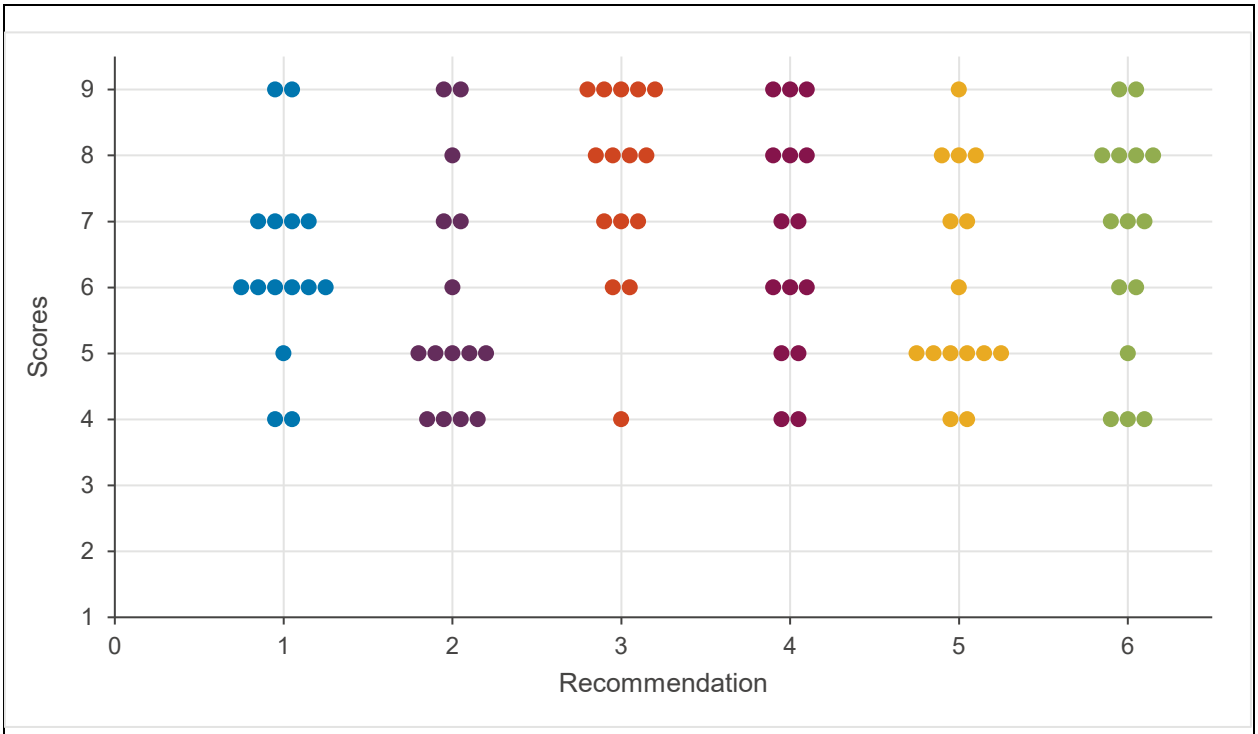

Figure 1: Distribution of scores awarded by the participants to each proposed RIVER recommendation (*n* = 15 per recommendation).

All six proposed recommendations received median scores of five or above. Three items received median scores indicating they were considered ‘critical’ by the working group, with the other three receiving median scores indicating recommendations considered ‘important, but not critical’. None received a single individual score below four, and each received at least one score of nine. Because all six proposed recommendations were considered, at the very least, “important”, all six were taken forward and developed into the recommendations appearing in this paper.

4. Development of the explanations

To further develop each recommendation and its accompanying explanation, the working group was split into three balanced sub-groups. Each sub-group was allocated two recommendations, corresponding to one of the themes laid out in table 1 (i.e. experimental design, experimental procedures and materials, or data handling, accessibility and visualisation). The text of each recommendation and explanation was iteratively developed by the relevant sub-group via meetings and collaborative working on shared documents. All recommendations were then reviewed and further refined by the whole group.

S1_User_Testing: RIVER Recommendations user testing methods

1. Purpose

A user testing study will take place following the publication of the RIVER recommendations on a preprint server. The purpose of user testing is to gather feedback on the recommendations from potential users. The study will evaluate whether each recommendation (and its accompanying explanation) is clear, helpful and easily understood in practice by *in vitro* researchers. Findings from user testing will be used to revise the manuscript, alongside peer reviewers' comments and invited feedback from the scientific community. Ethical approval will be obtained from a UK-based university prior to the study commencing.

2. Participants

A total of 10-15 *in vitro* researchers, who are in the process of writing a manuscript or will shortly be doing so, will be recruited. This number is sufficient to identify 80-90% of the issues in usability testing [1]. Because there is likely to be attrition between recruitment of users and receipt of their manuscripts (e.g. participants may decide they aren't yet ready to write a manuscript), a higher number of participants will initially be recruited. Previous experience (for example, user testing of the revised ARRIVE guidelines [2]) suggests that recruiting 20-30 researchers who meet eligibility criteria will be necessary for 10-15 to complete the process.

Participants from a variety of countries, research fields and career stages, and who use diverse types of experimental model, will be recruited. Recruitment will take place using social media and through contacts and public channels available to the RIVER working group and NC3Rs. Participants who complete the user testing process will receive a reward, in the form of a shopping voucher (or similar incentive).

3. Process

Participants will be asked to consult the RIVER recommendations preprint while preparing their manuscript, and to report their experiments in line with the recommendations. They will be asked to complete a form, indicating the location in their manuscript that information related to each recommendation is reported, or explaining why that information was not included. They will also be asked to provide general feedback on the content, clarity and ease of use of the recommendations.

Each participant's manuscript will be anonymised and independently reviewed by two members of the RIVER working group, to record what information is reported (if any) for each recommendation, and identify any missing information.

Once a manuscript has been reviewed, a brief follow-up interview with each participant will be conducted. The purpose of this interview will be:

1. to gather further feedback on the content, clarity and ease of use of the recommendations.
2. to discuss any discrepancies between the authors' form and the working group's assessment of the manuscript, and discuss how the recommendations could be revised to help with reporting those pieces of information.

Feedback from user testing will be used to revise the recommendations, ensuring they are fit for purpose and well understood by their target audience.

4. Results

The following data will be collected and reported:

- The number of researchers who expressed an interest in, and who eventually participated in, the user testing process
- Demographic information on the participants, including their country of work, career level, primary research field, type of *in vitro* model(s) used, and whether they are a native English speaker.
- Information on how the participants used the guidelines, and their experience of doing so. For example (and among other points) this will include whether any parts of the recommendation and explanations were particularly helpful or unhelpful, whether they helped to identify sources of bias in participants' studies, and whether the recommendations prompted the participant(s) to add or remove any information from their manuscript.
- Any revisions made to the RIVER recommendations as a consequence of user testing feedback, and the rationale behind these.

References

1. Faulkner L (2003). Beyond the five-user assumption: benefits of increased sample sizes in usability testing. *Behav Res Methods Instrum Comput* 35(3):379-83 doi: 10.3758/bf03195514
2. Percie du Sert N *et al.* (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLoS Biol* 18(7):e3000410 doi: 10.1371/journal.pbio.3000410