scientific reports

Check for updates

OPEN

A call to action to address critical flaws and bias in laboratory animal experiments and preclinical research

Hugh G. G. Townsend[1]✉, Klaus Osterrieder[2], Murray D. Jelinski[3], Douglas W. Morck[4], Cheryl L. Waldner[3], William R. Cox[5], Volker Gerdts[6], Andrew A. Potter[7], Lorne A. Babiuk[7] & James C. Cross[8,9]

During the design of hypothesis-driven, comparative laboratory animal experiments, investigators must control for cage effects, ensure full blinding and full randomization while adhering to established experimental designs, notably variations of the Completely Randomized Design and the Randomized Block Designs. Failure to meet these criteria introduces partial or complete confounding by multiple known and unknown variables, resulting in biased outcome measures and rendering valid statistical analysis impossible. Our analysis of a stratified, random sample of comparative laboratory animal experiments conducted in North America and Europe and published in 2022, shows that as few as 0–2.5% utilized valid, unbiased experimental designs. The failure of investigators to adopt valid, unbiased study designs undermines scientific rigour, squanders resources and animal lives, and impedes the reliable translation of preclinical research findings to human and veterinary medicine. We propose practical, achievable solutions focused on enhancing the rigour and validity of study designs. This includes developing a specialized group of scientists with expertise in the design of laboratory animal experiments and data analysis, to ensure future studies are conducted with the highest scientific standards.

"*The use of animals in research, teaching and testing is acceptable only if it promises to contribute to the understanding of environmental principles or issues; fundamental biological principles; or development of knowledge that can reasonably be expected to benefit humans, animals or the environment.*" Canadian Council on Animal Care guidelines (1997)[1].

> "*No one would now dream of testing the response to a treatment by comparing two plots, one treated and the other un-treated.*" R. A. Fisher and J. Wishart (1930)[2].

All laboratory animal experiments comparing groups of differentially treated subjects are conducted in cages, with small numbers of animals assigned to each cage. Both external and internal factors impact cage environments and the animals within[3–12]. As well, each cage of animals has an individual phenotype[13] and phenotypic plasticity[14] (the property of organisms to produce distinct phenotypes in response to environmental variation). For these reasons, no cage of animals is expected to respond to a specific treatment in precisely the same way as any other cage. Therefore, the effect of cage on study outcomes must be addressed during the design,

[1]Department of Large Animal Clinical Sciences, Western College of Veterinary Medicine (Emeritus) and Vaccine and Infectious Disease Organization (Retired), University of Saskatchewan, Saskatoon, SK, Canada. [2]University of Veterinary Medicine Hannover, Foundation, Hannover, Germany. [3]Department of Large Animal Clinical Sciences, Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK, Canada. [4]Department of Biological Sciences, Faculty of Science, University of Calgary, Calgary, AB, Canada. [5]Amphoraxe Life Sciences, 4103 Parkway Dr, Vancouver, BC V6L 3C9, Canada. [6]Vaccine and Infectious Disease Organization and Department of Veterinary Microbiology, Western College of Veterinary Medicine, University of Saskatchewan, Saskatoon, SK, Canada. [7]Department of Veterinary Microbiology (Emeritus), Western College of Veterinary Medicine and Vaccine and Infectious Disease Organization (Retired), University of Saskatchewan, Saskatoon, SK, Canada. [8]Faculty of Veterinary Medicine (Emeritus), University of Calgary, Calgary, AB, Canada. [9]McEachran Institute, Nanoose Bay, BC, Canada. ✉email: hugh.townsend@usask.ca

analysis and interpretation of the results of every laboratory animal experiment[11,15,16]. In addition, to be valid, such experiments must be fully blinded and fully randomized and utilize the correct unit of analysis[6,7,12,17–21].

There are several classical designs, developed by R.A. Fisher in the 1920's, that are available for use in laboratory animal experiments[12,22–24]. The designs most applicable to laboratory animal experimentation are variations of the Completely Randomized Designs (CRD) and the Randomized Block Designs (RBD). To increase the homogeneity among experimental units and minimize data variance related to confounding variables, animals in Completely Randomized Design experiments are randomly assigned to cages, with all animals in each cage assigned to the same treatment. The correct unit of analysis is the cage, and the outcome of interest is the average or weighted average[12,25] response of the animals within each cage[11]. The sample size for these experiments is equal to the number of cages assigned to each treatment, not the number of individual animals in each treatment group. The two-sample t-test and the analysis of variance (ANOVA) are appropriate for the statistical analysis. For variables assessed repeatedly over time, a paired t-test or repeated measures ANOVA is required. The results of the ANOVA's are amenable to full analysis of their residuals.

Although the Completely Randomized Designs provide a straightforward method for controlling cage effect, they generally result in increased variability in outcome estimates because each estimate includes variability from both cage and treatment effects[11]. Moreover, assigning treatments to entire cages of animals, rather than to individual animals, increases study costs due to the necessity for more cages of animals[25]. It may also raise ethical concerns if only one animal is assigned to each cage. This is a particular concern when using social animals like mice. To mitigate these issues, researchers should, whenever possible, utilize variations of the Randomized Block Designs.

Amongst the Randomized Block Designs, the Randomized Complete Block Design (RCBD) is particularly applicable to laboratory animal experiments. In its simplest form, this design controls for cage effect by assigning one animal from each treatment group to each cage, making each cage a block and the individual animal the unit of analysis. In this non-replicated design, where each treatment is assigned to exactly one experimental unit within each block, two-way ANOVA is the appropriate method of analysis, with treatment and cage being the factors of interest. Importantly, ethical restrictions limit the number of animals per cage to as few as five, thus limiting the number of treatment groups in the experiment to five as well.

The examination of the assumptions of ANOVA in non-replicated Randomized Complete Block Designs is confined to the generation of residual plots, homoscedasticity plots, and Q-Q plots. This limitation can be addressed by incorporating repeated measures into the experimental design, by replicating the design, or by utilizing Randomized Complete Block Designs with a split-plot structure[12,25,26]. However, implementing replications and split-plot designs requires the use of statistical models such as Mixed Models, General Linear Models and General Estimating Equations, that are not included in GraphPad Prism, the statistical package which appeared to be the program most commonly used in the publications that we reviewed.

The most common study designs utilized in laboratory animal experiments are the biased, Cage-Confounded Designs (CCD). These flawed designs are the result of assigning treatments to entire cages of animals and then performing the statistical analysis under the erroneous assumption that the animal is the correct unit of analysis[6,12,24,25]. This misidentification of the correct unit of analysis violates the fundamental assumption of data independence requirement for ANOVA. As a result, sample sizes are spuriously inflated due to data pseudoreplication, variances of outcome measures are decreased, confidence limits are narrowed, p-values are reduced and the probability of false positive results is increased[12,23,27,28].

A common and fatally flawed variation of the Cage-Confounded Design occurs when each treatment group is assigned to a single cage of animals. In this design, treatment effects become completely confounded by cage effects. Therefore, any observed differences among the outcomes of interest may stem from either treatment effects, cage effects or some combination of the two. Therefore, the variance attributable to treatment cannot be isolated from that attributable to the cage environment. In terms of ANOVA, the effective sample size is one ($n=1$), and the within-treatment (denominator) degrees of freedom equals zero ($F_{(n-1,\ 0)}$). Therefore, a valid ANOVA cannot be performed.

A primary objective of this paper is to help biomedical scientists understand that a necessary requirement for preventing bias in laboratory animal experiments involves a combination of controlling for cage effects, implementing full investigator blinding, and fully randomizing study procedures. These elements must be addressed during both the design and analysis phases of all comparative laboratory animal experiments. Without achieving these objectives, study designs will inevitably produce results that are biased to some degree, an outcome that cannot be resolved by any other means[12,29]. Ongoing failures to address these issues severely limit the repeatability, replicability, and reproducibility of laboratory animal experiments[12,30,31] as well as the potential for translating study results into meaningful applications within human and veterinary medicine[32,33].

The design and analysis of a randomized complete block design study

The importance of effective randomization and blinding in the prevention of biased results is well recognized and has been addressed repeatedly in publications that deal with the design of laboratory experiments[17,18,29,34–37] particularly since Kilkenny's 2009 publication on the quality of experimental design[19]. On the other hand, the critical problems that arise from failure to control for cage effect, employ the correct unit of analysis and adhere to Fisher's classical experimental designs have received little attention and are almost never mentioned in the publication of laboratory animal experiments. The results of comparative, laboratory animal experiments and related conclusions will only become consistent and reliable when the biomedical research community fully acknowledges and addresses the fundamental importance of these issues. Although these principles are reasonably straightforward, there are some nuances and complexities to address. We find it most effective to do so through the description of an experiment that exemplifies the principals involved. For this purpose, we present an example of a vaccination and challenge infection study that utilized a Randomized Complete Block Design

without replication. In this study, 16 five-to six-week-old, male Syrian hamsters, were purchased from Charles River Laboratories. All work with infectious SARS-CoV-2 was performed in a containment level 3 facility at the Vaccine and Infectious Disease Organization. All methods were performed in accordance with the relevant guidelines and regulations specified by the University of Saskatchewan's University Animal Care Committee (UACC) and the Animal Research Ethics Board (AREB) approved the animal work as per the guidelines of the Canadian Council of Animal Care's (CCAC) criteria (approval number AUP# 20200016 MOD5).The experiment was designed, conducted and reported in near full compliance with the Essential 10 ARRIVE guidelines. The sample size for the study was based on the results of a series of previous experiments and substantial experience with the model.

On arrival, four animals were randomly assigned to each of four individually ventilated, polycarbonate cages (bCON™ Biocontainment System, Lab Products, Aberdeen MD) using a random numbers generator (Microsoft Excel) and individually identified by ear notch and a subcutaneous, microchip transponder (see below). Throughout the study, animals were maintained on soft cellulose bedding (Biofresh™, Animal Specialties and Provisions, Quakertown, PA). Cages were spot cleaned daily and with bedding replaced every 2 weeks. Shredded crinkle paper (Enviro-dri®, Shepherd Specialty Papers, Milford, NJ) was provided for nesting. Animals had ad libitum access to rodent chow (5001 - Laboratory Rodent Diet, LabDiet®, Richmond IN) and water delivered via disposable plastic bags with stainless-steel sipper tubes (Edstrom Sipper Sack®, Avidity Sciences, Waterford, WI). Environmental conditions were maintained at a temperature of 21–23 °C and a relative humidity of 40–60%, with a 12-hour light/dark cycle (lights on at 07:00 h).

After seven days of acclimatization, each treatment (PBS controls and three different vaccine formulations) was randomly assigned to one animal in each cage. The treatments were coded, and the investigators blinded to the identity of the treatment groups to which the animals were assigned until after completion of the statistical analysis. Animals were immunized with two doses of their assigned vaccine, four weeks apart, or mock immunized with PBS. Three weeks after the second immunization, animals were challenged intranasally with live SARS-CoV-2 virus (1×10^5 $TCID_{50}$) in each nostril. and then euthanized 5 days later using deep anesthesia in a chamber with isoflurane.

Throughout the experiment, animals were monitored once daily until clinical signs were noted, then twice daily for the remainder of the trial. Parameters assessed post-challenge included activity level, body weight, and temperature, which was measured from a subcutaneous transponder (TP-1000™, Bio Medic Data Systems Inc, Seaford DE). Criteria for humane endpoints included a weight loss greater than or equal to 20%, a change in subcutaneous body temperature greater than or equal to 3 °C, dyspnea and, hunched posture or any other behavioral indicators of pain or distress. If any of these criteria had been met, animals would have been immediately euthanized. However, no animals met early removal criteria, and no unexpected adverse events occurred during the study.

Five days post-challenge, animals were euthanized using deep anesthesia in a chamber with isoflurane. This was followed by terminal cardiac exsanguination and blood sample collection. The chest was then opened for immediate collection of lung tissues. Viral RNA concentrations in each of five different locations in the lungs harvested from each animal were determined using quantitative reverse transcription PCR (qRT-PCR). The results are presented in Fig. 1 and the raw data and statistical analyses, and residual plots are presented

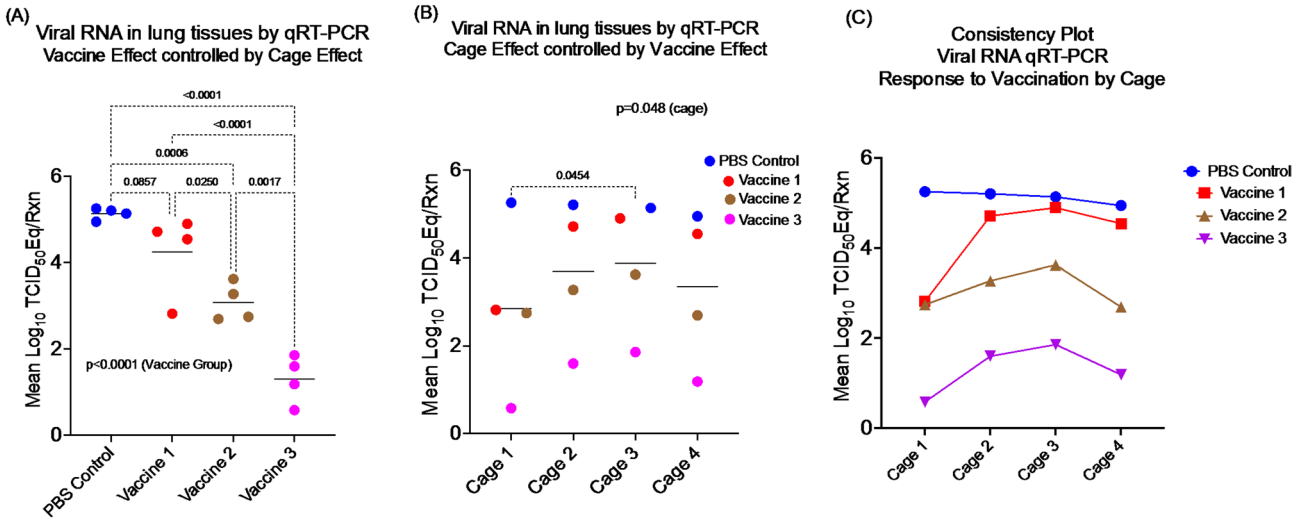

Fig. 1. A Randomized Complete Block Design challenge trial using 16 animals; four cages, four treatment groups, one animal per treatment group in each cage. **A** Two-way ANOVA showing significant differences in tissue virus concentrations among the vaccine groups ($p < 0.0001$) after controlling for the effect of Cage (horizontal bars denote mean values). **B** Two-way ANOVA showing significant differences among cages ($p = 0.048$) after controlling for the effect of vaccination due to a significant difference between cages 1 and 3 ($p = 0.0454$, horizontal bars denote mean values). (C). Consistency plot of the differences in virus concentration by cage and vaccine, showing the relatively constant differences within and between cages by vaccine.

in Supplementary Data #1. All animals consigned to the experiment completed the study and there are no missing data. Statistical analyses were performed using GraphPad Prism Version 9.5.1 (528), January 24, 2023. Prior to analysis, data were log-transformed to an approximate normal distribution and then an average virus concentration was calculated for each animal. The examination of the assumptions of ANOVA utilizing residual plots, homoscedasticity plots, and Q-Q plots are presented in Supplementary Data #1D. Visual assessment of interaction between vaccine group and cage is presented in Fig. 1C and Supplementary Data #1C. The experimental protocol was not pre-registered with an organization independent of the University of Saskatchewan.

To demonstrate the correct approach to the analysis of our example of a Randomized Complete Block Design experiment without repetition, we conducted a fixed-effect, two-way ANOVA[38] of the combined effects of vaccination status and cage assignment on the average concentration of virus in lung tissue (Fig. 1A). This analysis revealed a significant difference among the vaccine groups ($F_{3,9} = 52.42$, $p < 0.0001$) as well as a significant difference in the mean responses of the cages ($F_{3,9} = 3.96$, $p = 0.048$). Apart from the non-significant difference between the PBS control and Vaccine 1 ($p = 0.0857$), Tukey's Multiple Comparison Test revealed significant differences with respect to all other comparisons between the vaccine groups ($p \leq 0.025$). Figure 1B shows the results of Tukey's Test comparing the mean response among the cages ($F_{3,9} = 3.95$, $p = 0.048$) and a significant difference between Cages 1 and 3 ($p = 0.045$).

One of several advantages of the Randomized Complete Block Design is that it splits the data up into several smaller experiments which are then combined in the two-way ANOVA[23]. This makes it possible to generate the consistency plot of virus concentration by cage assignment (Fig. 1C). This plot shows that the order of magnitude, as well as the differences between the vaccines, were reasonably constant across the cages, and provides an indication of the internal validity and repeatability of the results of the experiment. To conclude, with this model, cage effect was controlled during both the design of the experiment and the analysis of the results. The associated graphs permit the visual assessment of the differences between the vaccination groups, the differences between the cages and the consistency of the results across the cages. Confounding by known and unknown variables was controlled through randomization of animals to both cages and treatments. Because the vaccine groups were commingled in each cage, the data were heterogenized[39–42] and therefore, provide some level of external validity, potential replicability and reproducibility of the experimental results.

In the years leading up to the start of the current study, we observed that the statistical analyses of most published, comparative laboratory animal experiments were flawed because the authors had employed Cage-Confounded Designs, with treatments assigned to cages while incorrectly using the individual animal as the unit of analysis. In every instance, the result of this error in design is confounding of treatment effects by cage effects, pseudoreplication of data[27] and some level of distortion of the results of the experiment. Further, in multiple instances, treatment groups consisted of fewer than six animals, suggesting that just one cage of animals was assigned to each group. This fundamental flaw in study design results in total confounding of treatment effects by cage effects, making it impossible to determine if the differences among the groups were due to treatment effects, cage effects or some combination of the two.

Limitations of the example data

To illustrate how a completely randomized block design experiment should be executed and analysed, we chose to use unpublished data from our existing files. There were some limitations regarding the control of bias in these data, but we do not believe this affected the utility of the data for use as an example study. We would have preferred to use published data for this purpose but were unable to find any appropriate examples. We could have used computer simulated data but concluded that readers might find this less convincing than data from an actual experiment.

With respect to limitations of the study design, although animals were randomly assigned to cages and then to treatments within cage, the order in which the treatments were administered within cage was not formally randomized. Also, within cage, the order in which lung tissues were sampled were not formally randomized. Although we do not believe that these deficiencies adversely impact the utility of the data as used in our example calculations, they should have been addressed in the study design. A formal sample size calculation for the experiment was not conducted. However, the example data comes from a large set of experiments conducted to meet government regulatory requirements relative to the development of a vaccine. The results of experiments carried out early in this series provided clear evidence as to the expected magnitude of clinically significant treatment effects, their expected variances and their statistical significance. Therefore, we had a reasonably precise and accurate idea as to an appropriate sample size. These estimates were likely more accurate than what could be achieved using standard sample size formulas and calculations.

Current practices in study design and data analysis of laboratory animal experiments

Systematic review of current publications

Having addressed the importance of controlling for cage effects, full randomization and blinding, as well as correctly identifying the unit of analysis, all in the context of classical experimental designs, we then focused on assessing the extent to which these objectives are achieved in published laboratory animal experiments. Specifically, we aimed to evaluate the prevalence of valid, unbiased experimental designs in the biomedical literature. To this end, we conducted a systematic review, following the PRISMA 2020 guidelines[43], of a random sample of research studies, stratified by country and published in 2022. Our investigation was confined to mammals. We defined a laboratory animal as a mouse, rat, hamster or ferret. Publications were sourced from PubMed, which we judged to be the most appropriate database for our purposes[44]. The country of origin for each publication was that reported in the animal ethics statement. Studies selected for detailed review and analysis

were those describing experiments comparing outcomes from two or more independent groups of animals (not animals from the same litter), each subjected to a different treatment or condition, with defined outcomes measured on one or more occasions during the experiment. We believe that the results of our random sample can be generalized to research published prior to and beyond 2022, provided they fit within the descriptors listed above.

The materials and methods of each publication were assessed for evidence of blinding of study personnel, randomization of procedures, correct identification of the unit of analysis and control of cage effects. Valid experiments were those that were fully randomized[18,20,45] (preferably with the aid of a random numbers generator) at every stage in the investigation where failure to do so might introduce systematic bias into the data (e.g., randomizing animals to cages, animals to treatments within cage, position of cages in their racks, submissions of samples to the laboratories and the statistical analysts). Full blinding[18,20,31] is achieved most efficiently through coding of animals and data or other recognized masking procedures. These procedures are aimed at concealing group assignments from all study personnel, preventing conscious or unconscious bias from influencing experimental conduct and measurements through to the completion of the analysis and interpretation of the results. Control of cage effects is achieved through the use of Completely Randomized Designs or Randomized Block Designs, with clear assurance that the experimental units are equal to the units of analysis[12] and the sample sizes are greater than one.

All studies included in the stratified random sample were conducted in Canada, France, Germany, the UK or the USA and published in 2022. Our initial search yielded 34,992 published papers. Employing a random sampling procedure normally used to detect disease in populations of animals[46] we estimated that to be 95% confident that we would detect at least one fully blinded, randomized experiment that was designed to control for cage effect, if the true prevalence of such studies was at least 3%, we should review 111 randomly selected publications. Arbitrarily, the sample size was then increased to 120 publications (USA = 60; Canada, UK, France or Germany = 60). Publications for review were selected by a single individual (HGGT). The strategy for doing so is described in the Supplementary Methods. All publications selected for full analysis were assessed using a stepwise, computer-aided search combined with manual scanning as described in the Supplemental Methods. Initially, each study was independently assessed by two investigators (HGGT and WRC) using separate spreadsheets (Microsoft Excel 2024). Any discrepancies were resolved through additional reviews and discussion. None of the original authors were contacted at any time for any purpose, including confirmation of their study designs.

In our stratified random sample of 120 publications from 2022, we did not identify a single instance of the use of a valid study design, 95% CI (0.0, 2.5). A summary of the results of this investigation is presented in Table 1 and all data and analyses are available in Supplementary Data #2. The prevalence of individual design features is as follows: full binding 2%, 95% CI (0.0,4.0); full randomization 0%, 95% CI (0.0, 2.5); no blinding or randomization of any nature, 42%, 95% CI (32.7, 50.0), and use of the correct unit of analysis in the statistical analyses, 8%, 95% CI (3.4, 13.3). As an indication of the lack of recognition of the importance of blinding and randomization in these experiments, only 38% of studies reported blinding on some level and only 43% reported some use of randomization. Some level of information about the methods used to blind investigators was reported in just 4 of the 43 partially blinded studies and in 3 of the 51 partially randomized experiments.

Design	Combined				Ca, Fr, Gr. UK			USA		
	Total	Count	Percent	95%CI	Total	Count	Percent	Total	Count	Percent
Full Blinding	2	120	2%	0.0, 4.0	2	60	3%	0	60	0%
Partial Blinding	43	120	36%	27.3, 44.4	22	60	37%	21	60	35%
Partial or Full Blinding	45	120	38%	28.8, 46.2	24	60	40%	21	60	35%
Methods of Blinding Described	4	45	9%	0.6, 17.2	4	24	17%	0	21	0%
Full Randomization	0	120	0%	0.0, 2.5	0	60	0%	0	60	0%
Partial Randomization	51	120	43%	33.7, 51.3	22	60	37%	29	60	48%
Partial or Full Randomization	51	120	43%	33.7, 51.3	22	60	37%	29	60	48%
Methods of randomization described	3	51	6%	0.0, 12.3	2	22	9%	1	29	3%
No Randomization or Blinding	50	120	42%	32.7, 50.0	28	60	47%	22	60	37%
Partial Randomization and Blinding	25	120	21%	13.6, 28.1	13	60	22%	12	60	20%
Randomization or Blinding	44	120	37%	28.0, 45.3	18	60	30%	26	60	43%
Correct Unit of analysis	10	120	8%	3.4, 13.3	8	60	13%	2	60	3%
Unbiased Study Design	0	120	0%	0, 2.5	0	60	0%	0	60	0%

Table 1. Categorization and distribution of the results of a random sample of 120 comparative laboratory animal studies, stratified by country and published in 2022. Sample sizes: USA = 60; Canada, France, Germany, UK = 60 (USA = 50%, Canada = 13%, France = 12%, Germany = 18%, UK = 8%), (mice = 80%, rats = 18%, hamster = 3%).

Specific reference to any of the classical experimental designs was not made in any publication. However, amongst the ten studies reporting the correct unit of analysis, nine housed individual animals in separate cages and thus, controlled for cage effect and met one of the basic requirements of a Completely Randomized Design. The remaining study met one of the requirements of a valid Randomized Complete Block Design and controlled for cage effect by commingling all treatment groups within each cage. But, due to incomplete blinding and randomization, none of these 10 studies had a valid, unbiased design. Therefore, in the remaining 110 studies (92%), the statistical analyses used an incorrect unit of analysis. As a result, regardless of any randomization or blinding, the findings were invalidated due to pseudoreplication of the data.

Limitations of the randomized, systematic review of the published literature

Our investigation was confined to 120 studies undertaken in a limited number of countries and species. We wished to concentrate our study on recent publications and therefore chose 2022, the most recent full year prior to the beginning of our investigation. We believe that our study results and conclusions can be generalized to other mammalian species, countries and journals, finding it unlikely that there is concentration of unbiased studies published in other journals or years or relative to other mamalian, comparative laboratory animal experiments.

It is possible that a very small number of authors may have carried out a fully blinded, fully randomized study, adherent to the requirements of a classical study design and controlled for cage effect, without having indicated that they did so. However, it is unlikely that this will have been the case in more than 3% of published studies.

Although we do not see it as a limitation, our study would have been strengthened and more informative if we had made specific note of the few studies employing a formal sample size calculation, as well as describing how and when this was done. By doing so, authors will have given some assurance to their readers that these calculations had been done a priori and correctly. Further, we should have noted if a primary hypothesis had been identified in any study that we reviewed as well as all unexplained imbalances in the number of animals in the study groups. Our assumption is that most laboratory animal experiments are designed with equal numbers of animals in each group. Authors rarely identified a primary hypothesis and frequently reported different numbers of animals among the groups, without explanation for having done so. Failure to declare a primary hypothesis, a priori, along with imbalanced treatment groups suggests that selective reporting of results is common in published laboratory animal experiments. Finally, we ought to have noted whether or not any investigators had registered their experiments as this limits the opportunity for selective reporting. From memory, none did. We recommend that investigators address these issues during study design, analysis and when reporting their results.

Study design in high impact journals

Numerous publications and guidelines, notably the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines[20] emphasize the critical importance of blinding[18], randomization, and understanding the concepts of the experimental unit and unit of analysis in the reporting of laboratory animal research. However, several recent studies provide compelling evidence that the development and dissemination of such guidelines have not resulted in substantial improvements in the reporting of animal experimental study designs[17,29,31,37,39,47–49]. Consequently, the findings of our systematic review of current literature were not unexpected. That said, it has been published that there are indications that the situation is improving, particularly in high-impact journals[34,48,50–53]. To investigate this finding from our own perspective, we analyzed 25 of the most recent mouse vaccination and challenge studies published March 30, 2024 or before, in four high-impact journals known for publishing substantial numbers of such studies: Nature Communications, NPJ Vaccines, Vaccines (Basel), and Frontiers in Immunology. Our rationale was that such studies are common, have a simple structure (generally a control group and one or more treatment groups followed over time), are easily blinded and randomized and are generally amenable to commingling of treatment groups. Publications were sourced using a PubMed search of each journal and sorted by the most recent date of publication with the last search performed on April 5, 2024 at 7:39 AM. Each publication was assessed using the same algorithm as employed in our stratified random sample. The results are presented in Table 2 and Supplementary Data #3. We show that in comparison to our stratified random sample of studies conducted in North America and Europe, the studies published in high ranked journals had numerically lower frequencies of all the attributes that we assessed, together with a greater frequency of failure to mention either blinding or randomization in the text of the publications. As well, amongst the publications from the high impact journals, only 2 studies approached a valid design through the utilization of one attribute of a Completely Randomized Design (animals were housed in individual cages) and no study employed a Randomized Complete Block Design. In summary, none of the sequence of 100 studies, published in the four high impact journals, utilized an unbiased study design.

Our findings clearly contradict previous reports suggesting that rates of blinding and randomization in laboratory animal studies are increasing[34,35,50,52–54]. This discrepancy arises because these publications focus on identifying any mention of randomization and/or blinding, whereas our focus is on the use of full implementation of both procedures[17,18,31]. We assert that reporting on partial or non-implementation of these practices does not ensure scientific rigor or provide a reasonable safeguard against bias. True and full randomization requires that all subjects or experimental units are randomized at every stage where failure to do so could introduce bias. Partial randomization inevitably leads to systematic errors, undermining the primary objective of ensuring treatment group comparability. Similarly, partial blinding is insufficient to eliminate bias due to preconceived notions or expectations, which can skew the interpretation of experimental results or the analyses. Partial blinding also introduces the risk of unjustified exclusions of data that conflict with expected outcomes, further compromising the validity of the results. Only full randomization and blinding can be expected to protect against the occurrence of biased research outcomes.

Design	COMBINED			NATURE COMMUNICATIONS			NPJ-VACCINES			VACCINES (BASEL)			FRONTIERS IN IMMUNOLOGY		
	TOTAL	COUNT	PERCENT	TOTAL	COUNT	PERCENT	TOTAL	COUNT	PERCENT	TOTAL	COUNT	PERCENT	TOTAL	COUNT	PERCENT
Full Blinding	0	100	0%	0	25	0%	0	25	0%	0	25	0%	0	25	0%
Partial Blinding	16	100	16%	6	25	24%	2	25	8%	5	25	20%	3	25	12%
methods of blinding described	1	16	6%	1	6	17%	0	2	0%	0	5	0%	0	3	0%
Full Randomization	0	100	0%	0	25	0%	0	25	0%	0	25	0%	0	25	0%
Partial Randomization	26	100	26%	8	25	32%	6	25	24%	7	25	28%	5	25	20%
Methods of randomization	0	26	0%	0	8	0%	0	6	0%	0	7	0%	0	5	0%
No Randomizaation or Blinding	64	100	64%	14	25	56%	18	25	72%	14	25	56%	18	25	72%
Partial Blinding and Randomization	6	100	6%	3	25	12%	1	25	4%	1	25	4%	1	25	4%
Partial Blinding or Randomization	30	100	30%	8	25	32%	6	25	24%	10	25	40%	6	25	24%
Correct Unit of Analysis	2	100	2%	2	25	8%	0	25	0%	0	25	0%	0	25	0%
Unbiased study design	0	100	0%	0	25	0%	0	25	0%	0	25	0%	0	25	0%

Table 2. Design and distribution of 100 recently published mouse vaccination and challenge studies published in four high-impact journals.

PARTIAL BLINDING		MANUSCRIPT	
		YES	NO
SUMMARY REPORTS	YES	6 (12.5%)	9 (18.8%)
	NO	3 (6.3%)	30 (62.5%)
PARTIAL RANDOMIZATION		MANUSCRIPT	
		YES	NO
SUMMARY REPORTS	YES	7 (14.6%)	20 (41.7%)
	NO	5 (10.4%)	16 (33.3%)

Table 3. Comparison of partial blinding and partial randomization in reporting summaries and the corresponding manuscripts (reporting "NO" or failing to report were both classified as "NO"). Comparison shows lack of agreement between the reporting summaries and the manuscripts.

Included in the instructions to authors submitting scientific reports to the Nature portfolio of journals is a directive that for experiments involving live vertebrates and/or higher invertebrates, the Methods section must include a statement that the authors complied with the ARRIVE guidelines, of which several were a focus of interest for our paper. In addition, it is required that authors "complete a Reporting Summary in which authors must provide a written disclosure on the use of randomization and blinding, even when the disclosure is negative". Out of interest, we elected to review and compare these reports with the manuscripts published in two Nature Family of Journals that require that authors complete a Summary Report. The results of this effort revealed a substantial degree of discord between the Manuscripts and the Reporting Summaries. The results of this investigation are summarized in Table 3 and Supplementary Data #3 and show limited agreement between the text of the manuscript and the Summary Reports.

Discussion

The problem

Our systematic review, which analyzed a stratified random sample of 120 journal submissions reporting on experiments conducted in Canada, France, Germany, the UK and the USA and published in 2022, along with a separate assessment of a series of 100 recent articles in highly ranked journals, did not identify a single comparative laboratory animal experiment that employed an unbiased design. Based on our representative sample of the current literature, we estimate that at least 97% of all comparative laboratory animal studies fail to adhere to valid experimental designs, lack adequate blinding and randomization, and fail to both control for cage effects and use the correct units of analysis (Cage-Confounding). Consequently, the outcome data are not amenable to valid statistical analysis. With respect to their design, such studies are without scientific rigor and cannot be expected to yield repeatable results.

The costs of failed experimental design are enormous

It is important that the research community appreciate that world-wide, more than 100 million laboratory animals are used each year in research and research related activities[55–57] at a cost in the billions of dollars[12,58]. It should not be accepted that greater than 97% of the comparative experiments to which laboratory animals are consigned are based on invalid study designs. This enormous waste of animals and research resources is unethical[32,58–62]. Equally serious, there are immeasurable costs associated with making critical decisions based on biased results, including the abandonment of valid lines of inquiry and the pursuit of false discoveries[58–62]. The consequences of using flawed study designs in laboratory animal experiments have serious implications for human health and safety since demonstrating safety and efficacy in laboratory animals is integral to the process of approving the initiation of human trials[33,39,63,64]. The generation and acceptance of flawed animal data in this process surely results in failed assessments of efficacy in Phase II/III human trials and, in the extreme, harm to human subjects.

How did we get here?

The practice of employing Cage-Confounded Designs in which treatments are assigned to entire cages with the outcome assessments based on the individual animals within cages has existed since the beginning of modern laboratory animal experimentation. Prior to the discovery of the classical experimental designs and their statistical analysis, this approach may have been adopted out of simplicity and on the assumption that the effects observed in one cage could represent the broader population, with the importance of confounding cage effects going unrecognized. However, ever since Fisher devised the classical randomized designs in the early 20th century, along with his development of the analysis of variance and its assumption of independence of data, the science community has known how to design valid animal experiments. These classical designs have been in use by agricultural crop and animal scientists for many years but not by preclinical investigators. In addition, based on our broad sample of the literature, only 21%, 95% CI (14, 28) of comparative laboratory animal experiments even mentioned both randomization and blinding. Within high-impact journals, our estimate is 6%, 95% CI (2,13). In both cases, these estimates should have approached 100%. In addition, only 5% (12/220) of all reviewed studies utilized the correct unit of analysis. In the remaining 95% of the publications, the assumption of independence was violated with pseudoreplication present in every instance. Therefore, the statistical analyses of the outcome data produced p values of falsely increased significance. As a consequence, the probability of false-positive results was also increased.

Peer review and check-list effectiveness

In principle, animal research studies undergo review for scientific merit and design at three key stages: research proposal review by funding agencies, which includes peer and administrative evaluations; experimental review by institutional or governmental animal care committees, guided by national regulations; and ultimately, peer review prior to publication in scientific journals. Clearly, this process has not yielded acceptable numbers of publications of preclinical studies based on unbiased designs.

Numerous high-impact journals, including several from the Nature Publishing Group[48,50,51,53,54,65] have highlighted the inadequate quality of reporting in laboratory animal research. Despite several focused initiatives, such as the publication of the ARRIVE guidelines in 2010 and their update in 2020, as well as the introduction of Nature's Reporting Summaries in 2013[65], these efforts have seen limited success. The effectiveness of the Reporting Summaries is limited because they address only a few bias-related factors, provide minimal information about these factors, and exhibit only moderate consistency between the claims made in the Reporting Summary and what is presented in the text of the publications regarding randomization and blinding (see Table 3; Supplementary Data #3). As well, it is apparent that merely reporting, without explanation or justification, that neither randomization nor blinding were employed, should be accepted as a valid contribution to bias control. These observations reflect a disinterest among authors, reviewers, and editors in the implementation and thorough reporting of blinding and randomization in laboratory animal studies. Furthermore, a significant limitation of all the reporting guidelines, checklists, and Summary Reports that we reviewed, is their focus on completed work rather than emphasizing strategies to minimize bias during study design, before data collection begins. Clearly, unless checklists are utilized during the design phase of a study, effective reduction in the prevalence of bias in research outcomes cannot be expected.

A call for action

Failures in study design are considered a root cause of inadequate repeatability, replicability, reproducibility, and translatability of preclinical research to human applications. We propose two steps aimed at addressing these challenges. Firstly, we urge every laboratory animal research unit, regardless of size, to include "rigour and validity in animal research" as a regular item for discussion. This action would increase awareness of issues related to study design and, over time, will profoundly enhance research quality. Such improvements will lead to more reproducible findings, minimize wasted resources, and improve the reliability of preclinical research, ultimately narrowing the preclinical-to-clinical translation gap. Research units committed to scientific rigour will gain credibility, become more competitive for funding and attract collaborations with like-minded researchers. However, change is unlikely to come easily as current practices are deeply ingrained in our research culture. For example, established investigators may resist new expectations, viewing them as unnecessary impediments to their work. They may regard scrutiny of study designs, such as calls for a priori declarations of primary hypotheses, planned approaches to statistical analysis or preregistration of research, as a challenge to their established expertise. Similarly, roadblocks related to entrenched beliefs and a lack of access to expert guidance are likely to arise within funding agencies, ethics boards, and institutional administrations. Nonetheless,

introducing regular discussions of rigour and validity of experimental designs into meetings of all stakeholders will provide a powerful mechanism for improving laboratory animal research quality.

Secondly, the most significant barrier to reducing bias in laboratory animal experiments is the inadequate education and training of preclinical scientists, with flawed understandings and practices being passed down from mentors to students[17,31,40,66–68]. To tackle this, we call for, and plan to undertake, the establishment of MSc and PhD programs focused on educating experts in preclinical research design and analysis, a course of action that is in line with recommendations that have been firmly stated by the National Institute of Health[62]. Graduates of these programs will possess the knowledge and skills to serve as educators, advisors, and consultants across the research landscape. Prospective candidates should come from diverse backgrounds, particularly animal science and veterinary medicine. Graduates will gain specialized knowledge of laboratory animal physiology, behavior, nutrition, medicine and surgery, along with advanced training in statistics and contemporary study design methodologies. Practical, in-depth exposure to research laboratories and animal facilities will further equip them to tackle everyday challenges in study design and execution. Many universities have the faculties, resources and collective knowledge to deliver these programs. Individuals with this training will not only serve as educators and consultants but will also conduct their own independent research aimed at learning more about rigorous and valid approaches to study design and data analysis. In addition to working with the research scientists, such individuals would find important positions in government and funding agencies, university administrations, ethics and editorial boards.

Along with many like-minded individuals, we assert that all stakeholders in laboratory animal research must actively collaborate to limit the failures inherent in the current designs of many experiments. We reiterate prior calls to action for stakeholders in preclinical research to take decisive measures within their circles of influence. Only through these concerted efforts can the biomedical research community improve reproducibility, uphold ethical standards in laboratory animal experiments, and build the confidence in preclinical research that is crucial for the health of both animals and humans.

Data availability

All data and statistical analyses are presented in the Supplementary Data Files.

Received: 5 April 2025; Accepted: 12 August 2025
Published online: 21 August 2025

References

1. Canadian Council on Animal Care guidelines. https://ccac.ca/Documents/Standards/Guidelines/Protocol_Review.pdf (1997).
2. Fisher, R. A. & Wishart, J. *The Arrangement of Field Experiments and the Statistical Reduction of the Results* (Imperial Bureau of Soil Science, 1930).
3. Deloris Alexander, A. et al. Quantitative PCR assays for mouse enteric flora reveal strain-dependent differences in composition that are influenced by the microenvironment. *Mamm. Genome*. **17**, 1093–1104 (2006).
4. Devor, M. et al. Sex-specific variability and a 'cage effect' independently mask a neuropathic pain quantitative trait locus detected in a whole genome scan. *Eur. J. Neurosci.* **26**, 681–688 (2007).
5. McCafferty, J. et al. Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. *ISME J.* **7**, 2116–2125 (2013).
6. Hooijmans, C. R. et al. SYRCLE's risk of bias tool for animal studies. *BMC Med. Res. Methodol* **14**. https://doi.org/10.1186/1471-2288-14-43 (2014).
7. Karp, N. A. & Fry, D. What is the optimum design for my animal experiment? *BMJ Open Sci.* **5**. https://doi.org/10.1136/bmjos-2020-100126 (2021).
8. Dunham, S. J. B. et al. Sex-specific associations between AD genotype and the Microbiome of human amyloid beta knock-in (hAβ-KI) mice. *Alzheimer's Dement.* **20**, 4935–4950 (2024).
9. Lemmens, V. et al. YF17D-vectored Ebola vaccine candidate protects mice against lethal surrogate Ebola and yellow fever virus challenge. *NPJ Vaccines* **8**. https://doi.org/10.1038/s41541-023-00699-7 (2023).
10. Streiff, C. et al. The impact of cage dividers on mouse aggression, dominance and hormone levels. *PLoS One* **19**. https://doi.org/10.1371/journal.pone.0297358 (2024).
11. Landes, R. D. How cage effects can hurt statistical analyses of completely randomized designs. *Lab. Anim.* **58**, 476–480 (2024).
12. Rowe, A. Recommendations to improve use and reporting of statistics in animal experiments. *Lab. Anim.* **57**, 224–235. https://doi.org/10.1177/00236772221140669 (2023).
13. Velasco-Galilea, M., Piles, M., Ramayo-Caldas, Y. & Sánchez, J. P. The value of gut microbiota to predict feed efficiency and growth of rabbits under different feeding regimes. *Sci. Rep.* **11**. https://doi.org/10.1038/s41598-021-99028-y (2021).
14. Helene Richter, S. Systematic heterogenization for better reproducibility in animal experimentation. *Lab. Anim.* **46**, 343–349. https://doi.org/10.1038/laban.1330 (2017).
15. Varholick, J. A. et al. Social dominance hierarchy type and rank contribute to phenotypic variation within cages of laboratory mice. *Sci. Rep.* **9**. https://doi.org/10.1038/s41598-019-49612-0 (2019).
16. Volianskis, R. et al. Cage effects on synaptic plasticity and its modulation in a mouse model of fragile X syndrome. *Philos. Trans. R. Soc. B Biol. Sci.* **379**. https://doi.org/10.1371/journal.pbio.3001873 (2024).
17. Landis, S. C. et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature* **490**, 187–191. https://doi.org/10.1038/nature11556 (2012).
18. Karp, N. A. et al. A qualitative study of the barriers to using blinding in in vivo experiments and suggestions for improvement. *PLoS Biol* **20**. https://doi.org/10.1371/journal.pbio.3001873 (2022).
19. Kilkenny, C. et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* **4**. https://doi.org/10.1371/journal.pone.0007824 (2009).
20. Percie du Sert, N. et al. Reporting animal research: explanation and elaboration for the ARRIVE guidelines 2.0. *PLoS Biol.* **18**. https://doi.org/10.1371/journal.pbio.3000411 (2020).
21. Festing, M. F. W. Design and statistical methods in studies using animal models of development. *ILAR J.* **47**, 5–14 (2006).
22. Festing, M. F. W. & Altman, D. G. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J.* **43**, 244–258 (2002).
23. Festing, M. F. W. Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *ILAR J.* **55**, 472–476 (2014).

24. Festing, M. F. W. The completely randomised and the randomised block are the only experimental designs suitable for widespread use in pre-clinical research. *Sci Rep* **10**. https://doi.org/10.1038/s41598-020-74538-3 (2020).
25. Walker, M. et al. Mixed-strain housing for female C57BL/6, DBA/2, and balb/c mice: validating a split-plot design that promotes refinement and reduction study design. *BMC Med. Res. Methodol* **16**. https://doi.org/10.1186/s12874-016-0113-7 (2016).
26. Altman, N. & Krzywinski, M. Points of significance: split plot design. *Nat. Methods*. **12**, 165–166 (2015). https://doi.org/10.1038/nmeth.3293 Preprint at.
27. Hurlbert, S. H. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**, 187–211 (1984).
28. Lazic, S. E. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.* **11**, 5 (2010).
29. Nunamaker, E. A. & Reynolds, P. S. 'Invisible actors'—How poor methodology reporting compromises mouse models of oncology: a cross-sectional survey. *PLoS One* **17**. https://doi.org/10.1371/journal.pone.0274738 (2022).
30. Plesser, H. E. Reproducibility vs. Replicability: A brief history of a confused terminology. *Front Neuroinform.* **11**. https://doi.org/10.3389/fninf.2017.00076 (2018).
31. Wilson, E. et al. Designing, conducting, and reporting reproducible animal experiments. *J. Endocrinol.* **258**. https://doi.org/10.1530/JOE-22-0330 (2023).
32. Reynolds, P. S. Between two stools: preclinical research, reproducibility, and statistical design of experiments. *BMC Res. Notes* **15**. https://doi.org/10.1186/s13104-022-05965-w (2022).
33. Errington, T. M. Building reproducible bridges to cross the valley of death. *J. Clin. Investig.* **134**. https://doi.org/10.1172/JCI177383 (2024).
34. Macleod, M. R. et al. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biol.* **13**, 1–12 (2015).
35. Leung, V., Rousseau-Blass, F., Beauchamp, G. & Pang, D. S. J. ARRIVE has not arrived: support for the ARRIVE (Animal research: reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLoS One*. **13**, e0197882. https://doi.org/10.1371/journal.pone.0197882 (2018).
36. Menke, J., Roelandse, M., Ozyurt, B., Martone, M. & Bandrowski, A. The rigor and transparency index quality metric for assessing biological and medical science methods. *iScience* **23**. https://doi.org/10.1016/j.isci.2020.101698 (2020).
37. Bergen, P., Munro, B. A. & Pang, D. S. J. Quality of reporting of prospective in vivo and ex vivo studies published in the journal of veterinary emergency and critical care over a 10-year period (2009–2019). *J. Vet. Emerg. Crit. Care*. **33**, 435–441 (2023).
38. Dixon, P. Should blocks be fixed or random? *Conf. Appl. Stat. Agric.* https://doi.org/10.4148/2475-7772.1474 (2016).
39. Dirnagl, U., Duda, G. N., Grainger, D. W., Reinke, P. & Roubenoff, R. Reproducibility, relevance and reliability as barriers to efficient and credible biomedical technology translation. *Adv. Drug Deliv. Rev.* **182**. https://doi.org/10.1016/j.addr.2022.114118 (2022).
40. Bailoo, J. D., Reichlin, T. S. & Würbel, H. Refinement of experimental design and conduct in laboratory animal research. *ILAR J.* **55**, 383–391 (2014).
41. Voelkl, B., Würbel, H., Krzywinski, M. & Altman, N. The standardization fallacy. *Nat. Methods*. **18**, 5–7 (2021).
42. von Kortzfleisch, V. T. & Richter, S. H. Systematic heterogenization revisited: increasing variation in animal experiments to improve reproducibility? *J. Neurosci. Methods* **401**. https://doi.org/10.1016/j.jneumeth.2023.109992 (2024).
43. Page, M. J. et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **372**. https://doi.org/10.1136/bmj.n71 (2021).
44. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G. Comparison of pubmed, scopus, web of science, and Google scholar: strengths and weaknesses. *FASEB J.* **22**, 338–342 (2008).
45. Lin, Y., Zhu, M. & Su, Z. The pursuit of balance: an overview of covariate-adaptive randomization techniques in clinical trials. *Contemp. Clin. Trials*. **45**, 21–25 (2015).
46. Prattley, D. J., Cannon, R. M., Wilesmith, J. W., Morris, R. S. & Stevenson, M. A. A model (BSurvE) for estimating the prevalence of bovine spongiform encephalopathy in a National herd. *Prev. Vet. Med.* **80**, 330–343 (2007).
47. Williams, J. L. et al. Weaknesses in experimental design and reporting decrease the likelihood of reproducibility and generalization of recent cardiovascular research. *Cureus* **14**, e21086. https://doi.org/10.7759/cureus.21086 (2022).
48. Flitti, D., Pandis, N. & Seehra, J. Still to ARRIVE at adequate reporting of orthodontic studies involving animal models. *Eur J. Orthod* **46**. https://doi.org/10.1093/ejo/cjae032 (2024).
49. Song, J. et al. A. Twelve years after the ARRIVE guidelines: Animal research has not yet arrived at high standards. *Lab. Anim.* **58**, 109–115. https://doi.org/10.1177/00236772231181658 (2024).
50. Han, S. H. et al. A checklist is associated with increased quality of reporting preclinical biomedical research: a systematic review. *PLoS One* **12**. https://doi.org/10.1371/journal.pone.0183591 (2017).
51. Checklists work to improve science. (2018). https://doi.org/10.6084/m9.figshare.6139937
52. Kousholt, B. S. et al. Reporting quality in preclinical animal experimental research in 2009 and 2018: A nationwide systematic investigation. *PLoS One* **17**. https://doi.org/10.1371/journal.pone.0275962 (2022).
53. Macleod, M. Did a change in nature journals' editorial policy for life sciences research improve reporting? *BMJ Open. Sci.* **3**. https://doi.org/10.1136/bmjos-2017-000035 (2019).
54. Cramond, F. et al. Protocol for a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. *Scientometrics* **108**, 315–328. https://doi.org/10.1007/s11192-016-1964-8 (2016).
55. Taylor, K., Gordon, N., Langley, G. & Higgins, W. Estimates for worldwide laboratory animal use in 2005. *Altern. Lab. Anim.* **36**, 327–342 (2008).
56. Carbone, L. Estimating mouse and rat use in American laboratories by extrapolation from animal welfare Act-regulated species. *Sci. Rep.* **11**, 493 (2021).
57. Taylor, K. Trends in the use of animals and non-animal methods over the last 20 years. *ALTEX* **41**, 503–524 (2024).
58. Keen, J. Wasted money in United States biomedical and agricultural animal research. In *Human-Animal Studies* vol. 22244–272 (Brill Academic, 2019).
59. Freedman, L. P., Cockburn, I. M. & Simcoe, T. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, 1–9 (2015).
60. Begley, C. G. & Ioannidis, J. P. A. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* **116**, 116–126 (2015).
61. Hawkes, N. Poor quality animal studies cause clinical trials to follow false leads. *BMJ* **351**. https://doi.org/10.1136/bmj.h5453 (2015).
62. Wold, B. & Tabak, L. A. ACD working group on enhancing rigor, transparency, and translatability in animal research. https://acd.od.nih.gov/documents/presentations/06112021_ACD_WorkingGroup_FinalReport.pdf
63. Henderson, V. C., Kimmelman, J., Fergusson, D., Grimshaw, J. M. & Hackam, D. G. Threats to validity in the design and conduct of preclinical efficacy studies: A systematic review of guidelines for in vivo animal experiments. *PLoS Med* **10**. https://doi.org/10.1371/journal.pmed.1001489 (2013).
64. Bailey, J. It's time to review the three Rs, to make them more fit for purpose in the 21st Century. *Altern. Lab. Anim.* **52**, 155–165. https://doi.org/10.1177/02611929241241187 (2024).
65. Anon, A. Reducing our irreproducibility. *Nature* **496**, 398 (2013).
66. Baker, M. Is there a reproducibility crisis? *Nature* **533**, 454 (2016).
67. Frommlet, F. & Heinze, G. Experimental replications in animal trials. *Lab. Anim.* **55**, 65–75 (2021).
68. Cobey, K. D. et al. Biomedical researchers' perspectives on the reproducibility of research. *PLoS Biol.* **22**, e3002870 (2024).

Author contributions

Conceptualization—H.G.G.T., J.C.C., K.O, M.D.J., V.G., A.A.P., and L.A.B., Provision of example data—V.G., Literature Reviews—H.G.G.T. and W.R.C., Data Analysis—H.G.G.T., Writing original draft preparation, H.G.G.T., J.C.C., K.O., M.D.J., Writing—Review and Editing, H.G.G.T., J.C.C., W.R.C., K.O., M.D.J., D.W.M., C.L.W. V.G., A.A.P., and L.A.B. All authors have read and agreed to the published version of the manuscript.

Funding

Unfunded research.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15935-4.

Correspondence and requests for materials should be addressed to H.G.G.T.