Drude *et al.*
Translational Medicine Communications (2022) 7:24
https://doi.org/10.1186/s41231-022-00130-8

Translational Medicine Communications

REVIEW

Open Access

Planning preclinical confirmatory multicenter trials to strengthen translation from basic to clinical research – a multi-stakeholder workshop report

Natascha Ingrid Drude[1*], Lorena Martinez-Gamboa[1], Meggie Danziger[1], Anja Collazo[1], Silke Kniffert[1], Janine Wiebach[1,2], Gustav Nilsonne[1,3,4], Frank Konietschke[2], Sophie K. Piper[2], Samuel Pawel[5], Charlotte Micheloud[5], Leonhard Held[5], Florian Frommlet[6], Daniel Segelcke[7], Esther M. Pogatzki-Zahn[7], Bernhard Voelkl[8], Tim Friede[9], Edgar Brunner[9], Astrid Dempfle[10], Bernhard Haller[11], Marie Juliane Jung[12], Lars Björn Riecken[12], Hans-Georg Kuhn[13,14], Matthias Tenbusch[15], Lina Maria Serna Higuita[16], Edmond J. Remarque[17], Servan Luciano Grüninger-Egli[18], Katrin Manske[19], Sebastian Kobold[19,20], Marion Rivalan[21], Lisa Wedekind[22], Juliane C. Wilcke[23], Anne-Laure Boulesteix[23], Marcus W. Meinhardt[24,25], Rainer Spanagel[25], Simone Hettmer[26], Irene von Lüttichau[27], Carla Regina[28], Ulrich Dirnagl[1] and Ulf Toelch[1*]

Abstract

Clinical translation from bench to bedside often remains challenging even despite promising preclinical evidence. Among many drivers like biological complexity or poorly understood disease pathology, preclinical evidence often lacks desired robustness. Reasons include low sample sizes, selective reporting, publication bias, and consequently inflated effect sizes. In this context, there is growing consensus that confirmatory multicenter studies -by weeding out false positives- represent an important step in strengthening and generating preclinical evidence before moving on to clinical research. However, there is little guidance on what such a preclinical confirmatory study entails and when it should be conducted in the research trajectory. To close this gap, we organized a workshop to bring together statisticians, clinicians, preclinical scientists, and meta-researcher to discuss and develop recommendations that are solution-oriented and feasible for practitioners. Herein, we summarize and review current approaches and outline strategies that provide decision-critical guidance on when to start and subsequently how to plan a confirmatory study. We define a set of minimum criteria and strategies to strengthen validity before engaging in a confirmatory preclinical trial, including sample size considerations that take the inherent uncertainty of initial (exploratory) studies into account. Beyond this specific guidance, we highlight knowledge gaps that require further research and discuss the role of confirmatory studies in translational biomedical research. In conclusion, this workshop report highlights the

*Correspondence: natascha-ingrid.drude@bih-charite.de; ulf.toelch@bih-charite.de

[1] Berlin Institute of Health (BIH) at Charité, BIH QUEST Center for Responsible Research, Berlin, Germany
Full list of author information is available at the end of the article

need for close interaction and open and honest debate between statisticians, preclinical scientists, meta-researchers (that conduct research on research), and clinicians already at an early stage of a given preclinical research trajectory.

Keywords: Confirmatory preclinical studies, Confirmatory preclinical multicenter studies, Preclinical multicenter studies, Robust evidence, Reproducibility, Clinical translation, Translational biomedical research, Multi-stakeholder workshop

Background

The decision to start a clinical trial to investigate a new drug or medical device is informed by preclinical studies to evaluate efficacy and safety. Depending on the medicinal product, some types of testing like toxicology studies are regulated and mandatory before moving from bench to bedside; others are specific to the disease, drug, and/or (animal) model. Here, we focus on preclinical efficacy studies where fewer regulatory prescriptions apply. The ultimate goal of such studies is to make knowledge claims [1]. Articulated on different effect levels, these include for example the claim of a specific role for a protein in a physiological process, or that an intervention will cure or slow the progression of a disease. To arrive at a knowledge claim, preclinical studies are performed in a stepwise approach. Hypothesis-generating exploratory studies evolve along a continuum through within-lab replications to knowledge-claiming confirmation. During this process, investigators need to continuously re-evaluate premises and refine study designs to increase validity and reliability. This includes defining Go/No-Go criteria for further studies already in the early stages [2]. When it comes to detailed guidance for this transition process, information on planning, conducting, analyzing, and evaluating confirmatory studies in preclinical research is scarce. The need for such guidance is emphasized by recent initiatives investigating evidence from single studies, for example in cancer biology, that find a substantial number of experiments that do not replicate. That is, effect sizes are substantially lower than in the original study and results are no longer significant [3]. Whereas this is not unexpected, and science has the potential to self-correct, efficient strategies need to be devised to foster translation into the clinic and generate patient benefit. This includes the essential questions of when and how to conduct a confirmatory study.

To close this gap, biostatisticians, preclinical scientists, clinicians, and meta-researchers held a workshop to discuss the aforementioned issues for preclinical multicenter confirmatory studies (see Figure S1 for the composition of workshop participants). Whereas the collaborative conduct of a study by more than one independent study site using shared protocols is common practice in clinical trials, this is a rather recent approach in the preclinical context [4]. Herein, most participating researchers currently conduct confirmatory studies funded by the *German Federal Ministry of Education and Research* [5]. Importantly, investigators aim to confirm their own previous exploratory research findings and underlying knowledge claims in a preclinical multicenter setting. Generated evidence should inform decisions to start a clinical trial. To develop guidance for conducting confirmatory studies, we have reviewed and discussed current approaches to identify what strength of evidence is needed before engaging in a confirmatory study and how evidence generation can be optimized in a confirmatory study concerning the knowledge claim. In this report, we will present suggestions from a transdisciplinary perspective and highlight open questions and opportunities for further research.

Main text

Towards robust evidence

For the decision to proceed to confirmatory experiments, criteria need to be defined a priori. These criteria reflect the evidence gathered so far and address the necessarily high uncertainty and possible bias of exploratory experiments. To evaluate robustness of evidence, two factors are of main importance: reliability and validity. Reliability refers to the characteristics of a result that reflect the level of replicability measured for example by effect size precision or statistical significance. Importantly, a reliable experiment is not necessarily valid as results might be replicable and still not reflect the underlying postulated mechanism. For this, experiments also need sufficient validity to substantiate the knowledge claim. Here, we recommend minimum criteria for validity and reliability to support the decision to conduct a confirmatory study.

Minimum reliability and validity criteria

In exploratory studies, low sample sizes often threaten the reliability of results. Two factors contribute to this. First, significant results do not necessarily reflect the existence of a biologically relevant effect. Second, even if they do the estimated effect size will be an overestimation of the actual effect. To understand the first issue, one must look at a set of scientific hypotheses that are experimentally tested. Some of these will reflect an underlying biologically relevant effect whereas others do not. The probability to detect a relevant effect is closely correlated

with the sample size. Low sample sizes as frequently seen in preclinical experiments and with that low statistical power will have decreased detection rates for these relevant effects [6, 7]. Additionally and inherent to statistical test procedures, experiments also produce false positives, usually 5% of all cases in which a biologically relevant effect does not exist. This results in a dilution of the small number of identified relevant effects by several false positives. That is, a significant finding derived in a low sample size experiment is at an increased risk of not reflecting a true cause-effect relationship. The second effect caused by low sample sizes is inflation of effect sizes for significant results. This so-called winner's curse is elicited by the applied p-value filter wherein only large experimental effect sizes yield significant results in low-powered experiments [8]. That is, even if experiments detect relevant effects the effect estimate carries a risk of inflation.

Consequently, when deciding whether to conduct a confirmatory study, the inflation of effect sizes and limitations of the p-value [9] need to be considered. If uncertainty about effect estimates is still high, within-lab replications could be a viable way to substantiate exploratory findings (see section *Within-lab replications as a road to rigorous evidence*). Alternatively, and similar to clinical trials, investigators can a priori determine a smallest effect size of interest that reflects biological or clinical relevance to argue for a specific mechanism of action or to predict the efficacy of an intervention, respectively. Such a lower bound could be informed by published effect size distributions, discussion with clinicians about viable clinical effects, and/or available resources that will only allow for a certain minimal effect size to be detected [10]. This discussion should involve biostatisticians and biomedical researchers who need to set decision-critical a priori criteria (e.g. smallest effect within confidence interval (CI) of exploratory study estimate) for progression to the next phase of experiments.

Regarding validity, the minimum set of criteria [11, 12] spans mainly three domains; internal, external, and translational validity. A high degree of internal validity is necessary already in the early stages. This not only includes measures to reduce the risk of bias such as randomization [13] and blinding [14], but also the use of validated methods that measure outcomes with low bias and high accuracy [15] (Table 1). To promote generalizability of results beyond the single experiment, external validity needs to be increased for example by investigating or systematically introducing sources of variation through systematic heterogenization. This can be achieved by varying genetic and/or environmental conditions, for example, by testing immune-competent animal models instead of specific pathogen free (SPF) immunocompromised strains [16, 17] or by the introduction of environmental variation in a multicenter approach. To what extent this is necessary and feasible already in exploratory stages is an open question. Another powerful tool that adds to external validity is triangulation where different methods and approaches are combined to support the same claim. If different methods yield converging evidence, validity of generated evidence increases at the potential cost of adding complexity to a study design [18]. Additionally, within-lab replications potentially increase external validity (see section Within-lab replications as a road to rigorous evidence). As the ultimate goal of these experiments is clinical translation, factors that are diagnostic for the human case need to be considered and outcomes defined to facilitate interpretation in the clinical context (translational validity). Particularly, (animal) models should reflect targeted aspects of human disease and converging evidence from different methods and contexts. We also recommend investigating the bioavailability of the drug before or very early in the confirmatory stage, which ideally includes pharmacokinetics. Here, dose-finding experiments should be performed before a large multicenter confirmation to either start with a predefined dose, or at least narrow it down to a minimum range. Other factors are less concerting for the decision to continue with a confirmatory study. For example, testing clinically relevant biomarkers and route of administration can be part of complementary experiments in the confirmatory phase. Those complementary experiments might be exploratory or considered flanking experiments to strengthen the evidence.

Within-lab replications as a road to rigorous evidence

If the minimum criteria (as presented in Table 1) are not met with the first exploratory study, replication experiments potentially serve as a powerful validation tool before conducting a larger (multicenter) study. In this context, within-lab replications or also mini-experiments [23] with refined experimental design and improved internal as well as external (by considering batch effects) validity will be valuable. Moreover, refined animal models generate evidence to assess translational potential in this early-stage replication e.g., from a low complex cell line-based xenograft cancer mouse model to a patient-derived xenograft model [24]. Using material from varying donors (if available) patient-derived models might enable the evaluation of different responses by better mimicking the clinical heterogeneity of the disease [25, 26]. In this context, companion diagnostics might be used to quantify the accumulation of a molecule at the target site or assess a hormone or receptor status to predict treatment outcomes or stratify subjects [27, 28].

Exact within-lab replications might also be used to increase the reliability of the results via increased

Table 1 Minimum criteria that need to be fulfilled/considered before starting a preclinical confirmatory multicenter trial. Best practices are based on existing (reporting) guidelines and sketch the ideal situation. However, there can be practical limitation that hinder e.g., blinding or randomization

Criteria	Minimum requirement	Best Practice	Restrictions/ Considerations
Internal Validity			
Blinding *Concealment of group allocation from one or more investigator(s) involved in a preclinical study*	Blinded outcome assessment	Blinding of treatment allocation, experiment(s), outcome assessment and analyses	Experiments in which the treatment allocation is directly linked to an obvious phenotypic difference from the start of the experiment (e.g. genetically modified mice with different fur colors)
Randomization *Using chance methods to allocate subjects to intervention and/or treatment according to a clearly defined probability distribution*	Completely randomized [13]	Block design and stratification within known (not post-hoc) important predicting strata (like bodyweight)	Social transfer of e.g. pain may limit randomization options [19, 20]
Inclusion/Exclusion *Differentiate between animal attrition or drop-out and (data)* ***outlier*** *management*	Clearly **a priori** defined inclusion/exclusion criteria Reporting of drop-out rate and/or animal attrition If data points are removed, it must be performed **before unblinding** according to a pre-defined protocol	Report full datasets and report all excluded animals with reason	*Inclusion/exclusion criteria can be based on animal welfare (severity assessment and human endpoint), on scientific outcome (e.g. three times SD) or on characteristics of the model (genotype, phenotype, stage of disease)*
Outcome	Primary outcome needs to be clearly defined (measurement unit and time point) and disease relevant (as defined involving a clinician)	Primary and secondary outcomes are clearly defined	
Quality Management/ Assurance *Including standardization (and harmonization) of protocols*	Protocols /work instructions and/or standard operating procedures in place Measures to assure quality of methods and models are defined (e.g. baseline measures across laboratories)	Harmonization of protocols across laboratories prior to the multicenter study (identification of differences) Training of experimenters	Different regulatory requirements regarding animal welfare in multi-center studies performed across different legal jurisdictions
Claim specification	Knowledge claim specification	Preregistration including specification of hypotheses (knowledge claims) and criteria for acceptance/ rejection	preclinicaltrials.eu animalstudyregistry.org osf.io
Statistical methods	Need to be defined in advance (which methods are to be performed and which assumptions been made) including sample size calculation	Preregistration [21]; Registered reports [22]	*Reach out to statistical consultants if needed*
Reliability *Consistency in a measurement*	Sufficient number of animals to assess the clinically or biologically meaningful effect and its associated uncertainty to inform sample size calculations	Increase sample size via within-lab replication to estimate effect size with adequate precision	*Within-lab replication can happen in parallel or across time (preferred)*

Table 1 (continued)

Criteria	Minimum requirement	Best Practice	Restrictions/ Considerations
Translational Validity *Extent to which a scientific finding can be translated from preclinical to clinical (human) contexts*	Animal model is relevant for disease and **reflects some** of its **characteristics** Indicating context of relevance (diagnostic manuals and categorical criteria or transdiagnostic approaches) *Be aware of model limitations!*	Include clinically relevant biomarker(s) and/or diagnostics For medicinal product: biodistribution and/or bioavailability Animal model is highly relevant and carries many disease characteristics And/or perform experiment using different (animal or human cell based) models/ tissue with complementary characteristics (**Triangulation**)	Experiments focusing on e.g. mechanistic understanding that do not aim directly at clinical translation

sample size and/or increasing the number of (smaller) batches [29]. This will decrease outcome uncertainty and aid in sample size planning for confirmatory studies. Ethical constraints, e.g. regarding studies including large animals, potentially prohibit stand-alone exact replication experiments. However, a replication study might be integrated as a positive or negative control group into the experimental design of a new exploratory study.

Ideally, exploration and within-lab replication studies have the potential to reveal effect modifiers, confounders, and colliders. This may require adjustment of experimental design, for example by including an estimate of drop-out rate either due to the animal model or due to the intervention that affects sample size planning. Information on such covariates can then lead to a refinement of e.g. the randomization scheme if body weight is affecting the outcome of a study. In this example, to control for the variation in body weight, the experiment could be split up into smaller blocks and interventions would be randomized to experimental units within each weight block. It can also support the selection of Go/No-Go decision points before confirmation. Finally, the decision about the transition from exploration to confirmation needs to include all stakeholders including preclinical and clinical researchers as well as biostatisticians.

Engaging in a confirmatory multicenter study -reality check

Irrespective of the generated evidence from an exploratory study, feasibility needs to be evaluated to decide whether a multicenter decision-enabling experiment should be conducted. This evaluation includes practical constraints such as available resources (can increased animal numbers be handled?) or ethical approval (replication experiments as *area of tension* [30, 31]), and medical need. According to the animal welfare act and Directive 2010/63/EU of the European parliament [32], an animal experiment can only be justified if it generates new knowledge and if that knowledge outweighs the harm for the animals [33]. Thus, confirmatory studies need to go beyond exact replications and generate diagnostic (=decision enabling) evidence about a knowledge claim [30, 34, 35]. In general, exploratory studies provide only preliminary evidence. Building on such initial findings, confirmatory studies allow generalization beyond specific experiments gathering support for the underlying knowledge claim. For this, investigators need to ensure that validity and scientific rigor are preserved at a high level throughout the preclinical research trajectory (Fig. 1).

Optimization of evidence generation during confirmation

The goal of the (multicenter) confirmatory study is to support a knowledge claim and potentially inform the decision to move to the clinic. Again, a clear a priori definition of Go/No-Go decision points and clearly defined primary and secondary outcomes are indispensable. Other parts of the planning process are less generalizable (Fig. 2). Some of these aspects are beyond the scope of this manuscript and we will solely focus on biometry related issues or practical constrains/aspects (v-vii) (Fig. 2).

Protocols, standardization and systematic heterogenization

One important step in conducting multicenter studies is harmonization of protocols (Fig. 2 (i, v)). In this process, involved laboratories need to decide on which aspects of the experimental protocols need standardization and which will systematically vary between centers. Important aspects that need to be standardized and quality controlled include the treatment scheme to ensure comparable dosage and the same quality of the drug. Additionally, quality control measures identified through initial baseline studies are recommended. A comparison of outcomes from control groups for example can identify potential problems between centers early on. Knowledge about center variability and information on factors that influence variance of results can be gained by introducing systematic heterogenization. This includes comorbidities and the use of both sexes [36, 37]. The latter is considered a minimum requirement in a confirmatory approach except for sex-related diseases like prostate cancer or in case of well-grounded arguments.

Heterogeneity will also be introduced by each study center. One naturally occurring source of variation is the different experimenters themselves. However, the latest literature indicates that this is less of an issue, particularly if all involved parties are well trained [38, 39]. To assess replicability of results across centers, a low number of centers already is sufficient. A minimum of two participating laboratories may already be sufficient and the added value of additional laboratories decreases rapidly [37]. A small number of centers precludes, however, estimation of between center heterogeneity. Here, strategies need to ensure that centers actually can be jointly analyzed. Concerning animal experiments, husbandry conditions including food, temperature and cage mates will most likely vary between centers and laboratories and need to be considered if those affect the outcome [20].

Primary outcomes should be complemented by evidence from other sources. Here, selection of partner laboratories can also be based on such complementary methods and approaches. When developing drugs to

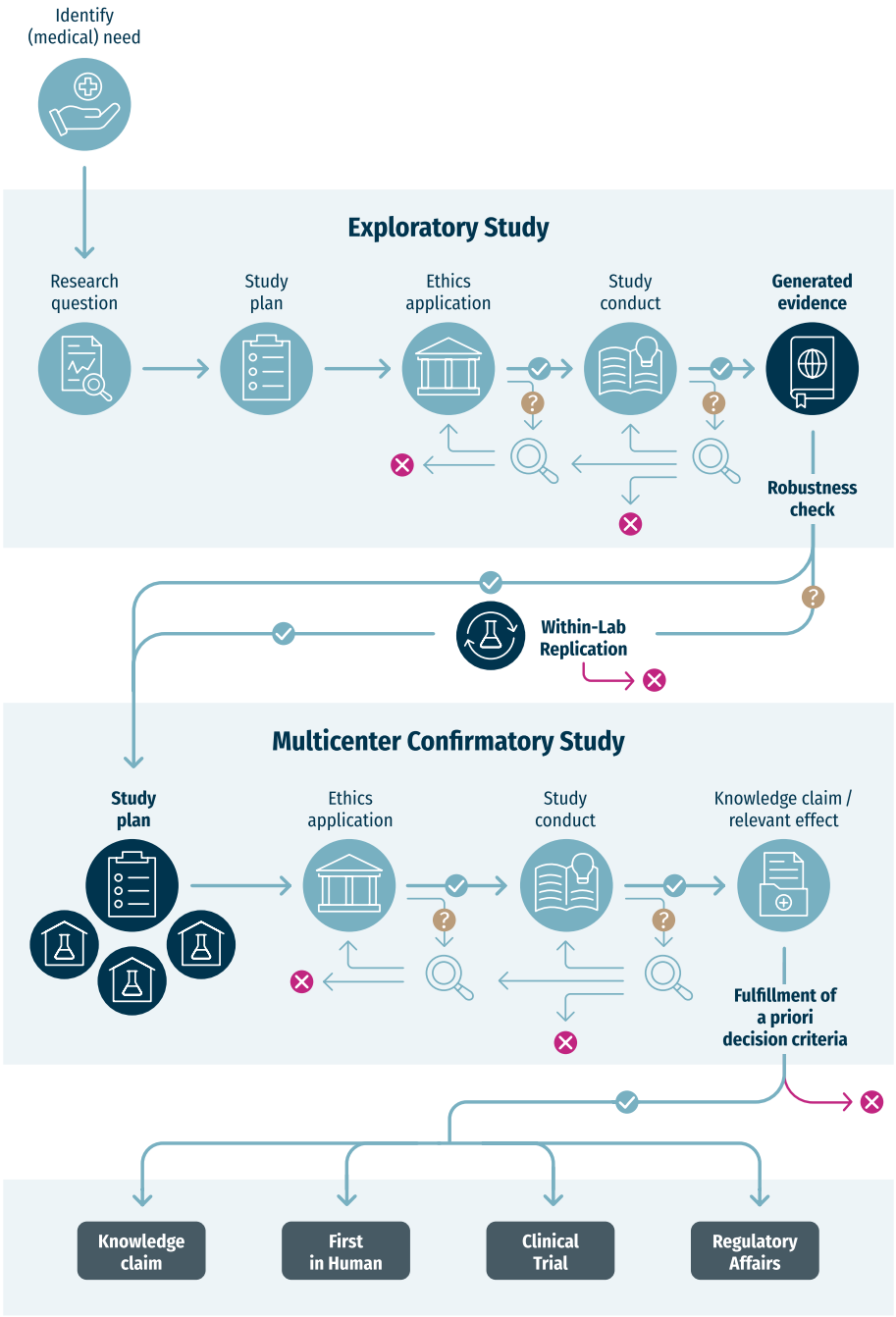

Fig. 1 Simplified illustration of a preclinical research trajectory, starting from exploration towards confirmation considering robust study design, minimum validity criteria to finally engage in a confirmatory multicenter study. At some steps a decision is required (loupe) whether to proceed with the study (check mark, blue), whether refinement of i.e., experimental design or (animal) model (question mark) is needed or whether (a priori defined) No-Go criteria were met to stop an experiment completely (red X). A robustness check after exploration should be used to decide if a within-lab replication is required before the multicenter confirmation. If minimum validity and reliability criteria are already met during exploration, a multicenter study might be planned without further in-house replication. Icons in dark blue (Generated evidence, within-lab replication, and study plan of the multicenter confirmatory study) highlight the focus areas of this review

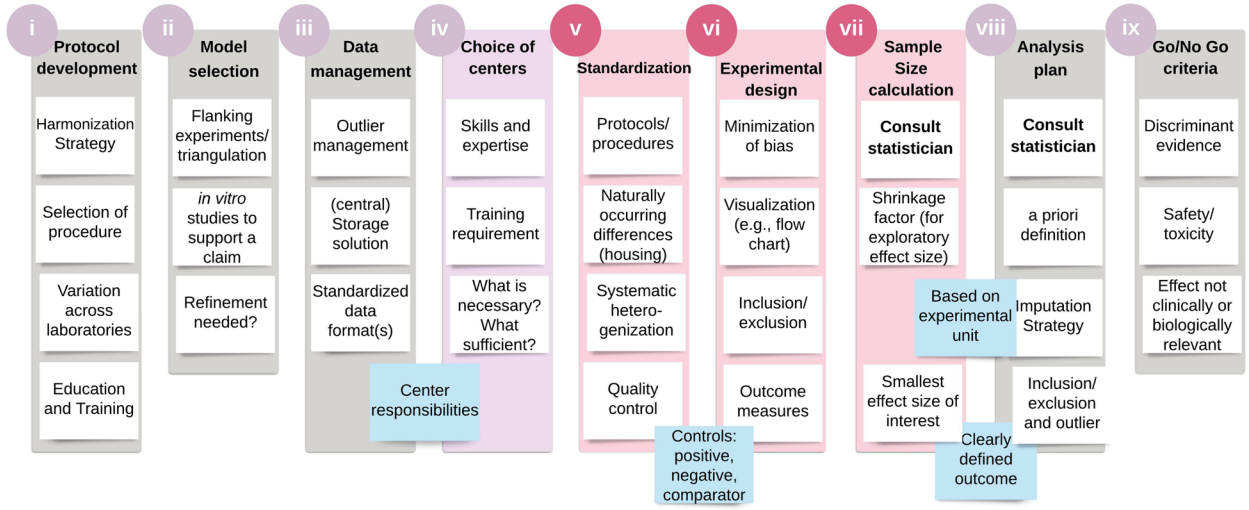

Fig. 2 Steps involved in planning a confirmatory study. Highlighted (pink) aspects (v – vii) are discussed in more detail. Whereas other aspects (grey, (I – iii and viii—ix) are equally important, they are beyond the scope of this manuscript and are solely mentioned in the context of experimental design, standardization and/or sample size calculation. Highlighted in purple (iv) is the choice of participating study centers. Here, skills and expertise as well as training requirements are important selection criteria. Blue notes cannot be assigned to only one aspect, but need to be considered at different stages, e.g., sample size calculation as well as the analysis should be based on the experimental unit

treat a symptom associated with multiple diseases, additional animal models can increase the external validity and predictability of translational success. Including but not limited to numerous existing animal models of neuropathic pain that can e.g., be chemotherapy-induced, emerge from cancer pain, or be mechanically introduced (sciatic nerve injury) [40, 41]. Another example are patient derived 3D cell cultures to gain a deeper understanding about underlying mechanisms and to capture effects only seen in human cells. By increasing the number of donors or models to support a research claim, the validity of an observed effect can be increased (triangulation [42]). For studies that aim at clinical translation, translational validity should be improved by including (several) biomarkers or other diagnostic tools [43, 44] in the analysis and/or experimental design. For drug efficacy testing, control groups in the confirmatory study should include a competitor drug i.e., clinical standard treatment and/or other negative and/or positive control groups. Researchers should be in close, early on contact with regulatory authorities to ensure that experiments already incorporate requirements for approval. To avoid increasing the sample size by additional positive and negative control groups, it can be feasible to consider historical cohorts [45, 46] or an unbalanced design [47, 48] with smaller but more control groups (multi-arm design) that can be pooled. The latter two points led to extensive discussions between the authors and should thus be viewed as controversial [49].

Sample size calculation for confirmatory studies

The basis for sample size calculation is the anticipated effect size that is defined in various ways [50, 51]. Herein, we refer to effect size as a mean difference divided by a measure of spread. In a typical preclinical efficacy study, that could be the difference between the mean of the primary outcome measure of an intervention group and of the control group divided by the pooled standard deviation [52]. As already mentioned earlier, the effect size estimate from exploratory studies tends to be inflated ("winners curse") [8]. Basing a sample size calculation of a confirmatory study on such an inflated effect size results in an underpowered study that runs the risk to miss an existing effect. This is aggravated in experiments with low internal validity [8, 53]. Sample size calculations for confirmatory studies should take this potential effect inflation into account and apply a shrinkage to exploratory effect size estimators to avoid underpowered studies. This also applies to effect sizes from published studies that are exploratory. This needs not necessarily be stated in the published study, but we recommend treating all research that does not explicitly state its confirmatory nature as exploratory. In case several prior studies are available (pilot, exploration, mini-experiments), effect sizes can be pooled via meta-analyses if heterogeneity between experiments is limited. Moreover, effect sizes do not typically extrapolate from animals to humans and are potentially smaller in humans [54]. It is thus necessary to apply shrinkage to effect sizes from exploratory studies, the exact magnitude, however, is still a matter of debate.

An alternative approach is to define a smallest effect size of interest as outlined above. This will set a lower bound under which results are no longer considered worthwhile exploring. Choosing such a threshold needs to reflect knowledge of the human disease, biology, effect size distribution in previous studies using similar model systems, available resources, and feasibility considerations [10]. That is, if the smallest effect size of interest is set too high the experiment will not be able to detect an actually existing effect. Contrary, an unnecessarily low smallest effect size of interest potentially requires a substantial number of resources and animals threatening the reduction principle of the 3R. In this context, in progressive diseases, clinicians can inform early treatment time points and evaluate how closely models reflect disease progression.

Once an effect size is chosen, this has an implication on the statistical power. With discussions on the utility of p-values and standard threshold of $p < 0.05$, the planning of a confirmatory trial can have a stricter bound such as a threshold of $p < 0.005$ or an increased power of for example 0.9 [55–57]. Again, this has to be weighed against the increased effort and cost–benefit calculations are necessary to avoid spending resources that could be used for other complementary studies [57, 58]. In confirmatory studies, strict correction for multiple comparisons should be applied to preserve the pre-specified false positive rate. As there is considerable uncertainty about the true effect, power could be calculated across a range of plausible effect sizes [59], instead of a point estimate to illustrate limitations for investigators. Particularly when confirmatory studies are conducted in a sequential manner [60], this may increase efficiency. Moreover, as the exploratory study has already registered the direction of the effect, sample size calculations and subsequent analysis can be based on one-sided tests. However, in case of an underpowered exploratory study aiming at mechanistic understanding confirming a prior knowledge claim, a sign error (type-S error) can occur where the replication detects an effect estimate in the opposite direction of the initial experiment or the actual effect size [61].

Multicenter considerations

A balanced design, where each center is allocated the same number of animals, is considered ideal as it increases the precision of estimates under between-center heterogeneity. One advantage over clinical trials here is that recruitment differences can be held to a minimum. Heterogeneity between centers is not due to different patient populations with different comorbidities but as outlined above most of the heterogeneity is systematically implemented in advance. The randomization to centers should take these previously planned factors into account in a block randomization scheme across centers. That is, factors need to be stratified and centers should for example test equal numbers of male and female animals, or animals from similar weight categories should be allocated to treatments similarly across centers. For this, a small number of additional animals may be needed to ensure a balanced design over all centers. Noteworthy, the impact on statistical efficiency with unequal or equal numbers of subjects in different centers also depends on the type of estimator used (e.g., fixed vs random effects). Finally, unbalanced numbers are not necessarily a sign of poor planning but a consequence of varying capacities or breeding of animals [62].

It is important to consider which experiments need to be performed by the initiating institute and which experiments by the partner laboratories. If a within-lab replication already indicated within-lab replicability of a result within the initiating institute, then this lab potentially does not need to perform the analogous experiment, but instead proceeds with triangulating evidence, a different strain, a different (large) animal model or flanking ex vivo experiments. In agreement with the initiating lab, partner labs can consider only selectively replicating core results to save on resources. Core results refer to assessment of the primary and important secondary outcome variables. If a costly method like single cell sequencing has been conducted in the initiating lab, a replication across all labs could lead to an undue increase in costs with little generation of additional insights. With respect to the animal model, subsequent designs are recommended (rodents—> non-rodents—> non-human primates). As sample sizes in large mammals including non-human primates typically need to be smaller due to ethical constrains, a smaller number of centers may be acceptable. It is an open question to which extent evidence from rodent experiments can be extrapolated to large animals and inform sample size planning. The effect size magnitude in rodents may neither translate to larger animals nor to the human case.

Reporting of confirmatory multicenter studies

Next to standard guidelines in preclinical research like ARRIVE [11], there are few points that are especially relevant when reporting a confirmatory study. This includes the provision of raw data to enable meta-analysis. Meta-analysis can help to cumulate evidence, find commonality, and develop guidance for best practices. In this context, it is crucial to transparently include and report **outliers** (data) as well as dropout rates (i.e., animal attrition). Standardization (or normalization) of all data

to one control group should be avoided. For better and transparent visualization of data, for example forest plots are suitable to show center specific data. With the a priori definition of No-Go decision points and potential failure of confirmation studies, it should be common practice to publish also null results.

Conclusion

Summarizing remarks and limitations

Even though confirmation studies are seen as essential part of preclinical research, so far little guidance exists on how to conduct such a confirmation. Here, we mapped out strategies to conduct such studies (see Table 2 for summary points and open questions). We acknowledge that no one size fits all; rather a broad set of recommendations applies that need to be adjusted to individual research fields and specific questions. Importantly, our recommendations are based on a scenario where an initial finding or exploratory study prompts the very same investigators to initiate a replication. This deviates from recent attempts where initial findings from other researchers were replicated on a larger scale [3, 63, 64]. These studies revealed that in many cases a replication could not even be attempted due to missing protocols or other aspects of reporting. In contrast, here we explore the scenario where researchers team up to confirm a knowledge claim. That is, confirmation in this case is not about an exact replication but rather to efficiently generate evidence to substantiate the knowledge claim and enable a decision to start a clinical trial.

Towards this goal, we described criteria to decide when to start a confirmation study, how to use within lab replications to arrive at or reinforce such evidence, and how to plan a multicenter study. This guidance is, however, based on the authors' and workshop participants' experiences and fields of expertise and is consequently focused on drug development and efficacy studies. While some of the ideas might be applicable for diagnostic and biomarker development, this is beyond the scope of this manuscript and requires further consideration.

Moreover, we have not addressed one important aspect in confirmatory multicenter studies. That is, has the confirmation been successful or not. Previous replication projects have shown there are numerous ways to define replication success [55, 63, 65, 66]. It is, however, unclear which of these criteria apply to confirmations and how they can guide decisions towards clinical trials.

As additional limitation, we foremost see that confirmatory projects are resources intense, and funders are less inclined to fund confirmatory research. Current developments with a funding line particularly for confirmatory studies in Germany and NIH initiatives [67] show that funding opportunities exist and probably will arise more

Table 2 Summary points and recommendations for the conduct of a confirmatory multicenter study including open questions that require further discussion and will be subject matter of future research

Summary points	Open Questions
Minimum validity and reliability criteria need to be fulfilled before engaging in a confirmatory multicenter study (Table 1)	Are dose–response effects a prerequisite for the confirmation?
If uncertainty is still high, optimization of evidence via (within-lab, in-house) replication studies to (i) increase sample size, (ii) improve internal validity, (iii) introduce systematic homogenization and/or (iv) flanking experiments	What if evidence from pilot, exploration and within-lab replication are **contradictory** (positive and negative results)?
(Standardized) protocols should be in place before starting a confirmatory study	-
(Animal) Model(s) should be disease relevant and limitations be acknowledged	-
Depending on the experimental objective control groups should include positive, negative controls and/or in case available a comparator from standard clinical care	What requirements need to be fulfilled to use historical control groups?
For planning a confirmatory study, sample size calculation should be based on smallest effect (size) of interest (clinical/biological relevant) or a shrinkage of the effect size(s) from exploratory studies should be considered	**Field specific effect sizes distributions** are scarce, how can the situation be improved? What is the **optimal approach** to calculate the sample size?
Flanking experiments (triangulation) might be performed early on and are highly recommended for confirmatory studies	How can in vitro studies be integrated in the confirmatory study design and sample size calculation?
Introduction of sources of variation like sex or strain (systematic heterogenization)	How to best **balance** standardization and systematic heterogenization?
Multicenter considerations include (i) harmonization of protocols, (ii) skills and expertise of partner lab(s), (iii) balanced design and (iv) block randomization across centers	Which experiments should be confirmed in several laboratories?

frequently in the future. With such funding also recognition for confirmatory research will grow in a similar way. Broader funding of specific confirmatory projects will open opportunities to reevaluate and refine the presented recommendations. Moreover, field specific strategies may evolve that will ultimately contribute to translation as a science with strong theory building at its core.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41231-022-00130-8.

Additional file 1: Figure S 1. Composition of workshop participants as per field of expertise. The organizing team mainly consists of meta-researchers with a clinical or preclinical background. Other refers to e.g., stakeholders from funding agencies that also attended the workshop. **Methods S1.** Multi-stakeholder workshop method-Workshop design. **Glossary.** Glossary of terms.

Acknowledgements
The authors would like to thank all workshop participants that were not part of the writing group for their valuable input and contributions.

Authors' contributions
ND, LMG, MD, AC, SK, JW, UD, and UT organized the workshop and round table discussion, guided discussions and summarized and clustered results. ND drafted the first version of the manuscript. All authors took active part in conceptualizing the manuscript (workshop, roundtable discussions, individual meeting, and written feedback). ND, MD, AC and UT wrote the glossary. SG, SKo, and EB revised and amended the glossary. All authors have substantively revised and/or edited the manuscript during the process. All authors have approved the submitted version of this manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. N.D., U.T., U.D., S.K., A.C., M.D.: This work was supported by the Federal Ministry of Education and Research (BMBF FKZ 01KC1901A ('DECIDE')).
R.S.: Financial support for this work was provided by the Ministry of Education and Research Baden-Württemberg for the 3R-Center Rhein-Neckar (www.3r-rn.de) and the Bundesministerium für Bildung und Forschung (BMBF) funded AhEAD consortium (01KC2004A).
E.M.P.-Z.: This work was supported by the Federal Ministry of Education and Research (BMBF, 01KC1903).
L.R.: This work was supported by the Federal Ministry of Education and Research (BMBF, 01KC1903A).
L.W.: BMBF 01KC2006A
SH, IvL, BH and CR would like to acknowledge funding by the German Ministery for Research and Education (BMBF 01KC2012A).
M.T. receive funding by BMBF; (Project "RSV Protect", 01KC2007A).
S.G.: Acknowledgement to the University of Zurich for funding my PhD.
S.Ko. is supported by the Marie-Sklodowska-Curie Program Training Network for Optimizing Adoptive T Cell Therapy of Cancer funded by the H2020 Program of the European Union (Grant 955,575, to S.K.); by the Hector Foundation (to S.Ko.); by the International Doctoral Program i-Target: Immunotargeting of Cancer funded by the Elite Network of Bavaria (to S.Ko. and S.E.); by Melanoma Research Alliance Grants 409,510 (to S.K.); by the Else Kröner-Fresenius-Stiftung (to S.Ko.); by the German Cancer Aid (to S.Ko.); by the Ernst-Jung-Stiftung (to S.Ko.); by the LMU Munich's Institutional Strategy LMUexcellent within the framework of the German Excellence Initiative (to S.E. and S.Ko.); by the Bundesministerium für Bildung und Forschung (S.Ko.); by the European Research Council Grant 756,017, ARMOR-T (to S.Ko.); by the German Research Foundation (DFG) (to S.Ko.); by the SFB-TRR 338/1 2021–452,881,907 (to S.Ko.); by the Fritz-Bender Foundation (to S.Ko.) and by the José-Carreras Foundation (to S.Ko.).
M.W.M.: Financial support for this work was provided by the Bundesministerium für Bildung und Forschung (BMBF) funded confirmatory call: AhEAD (FKZ: 01KC2004A), the MWK-funded 3R Center Rhine-Neckar (FKZ: 33–7533.-6–1522/9/4) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID: ME 5279/3–1.
L.H., C. M., and S.P. receive funding from the Swiss National Science Foundation (project number 189295, http://p3.snf.ch/Project-189295).
L.H. receive funding through BIH/QUEST Visiting Fellowship.
T.F. is grateful for support by the German Research Council DFG (FR 3070/4–1) and the German Center for Cardiovascular Research DZHK (81Z0300108).
F.K. is funded by the Deutsche Forschungsgemeinschaft DFG (KO 4680/4–1).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
ND is external consultant and animal welfare officer at Medizinisches Kompetenzzentrum |c/o HCx Consulting GmbH | Brandenburg, Germany.
S.Ko. has received honoraria from TCR2 Inc, Boston, GSK, BMS and Novartis. S.Ko. has received licensing fees from TCR2 Inc and Carina Biotech. S.Ko. has received research support from TCR2 Inc and Arcus Biosciences.
E.M.P.-Z.: During the last 3 years, E.M.P.-Z. received financial support from Grunenthal and Mundipharma for research activities and advisory and lecture fees from Grünenthal and Novartis. In addition, she receives scientific support from the German Research Foundation (DFG), the ERA-NET programm via the Federal Ministry of Education and Research (BMBF), the Federal Joint Committee (G-BA) and the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777500. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. All money went to the institutions (WWU/UKM) E.M.P.-Z. is working for (no personal fees).
S.G.: is a member of the cantonal animal research committee of the canton of Zurich, Switzerland (regulatory body), president of the scientific grassroots think tank ""Reatch! Research. Think. Change."" which—among other topics—advocates for responsible animal research. S.G. is a member of the board of the association ""Animal Research Tomorrow"" (formerly known as ""Basel Declaration Society"") See personal website for full disclosure of memberships and employments: https://www.servangrueninger.ch/offen-ehrlich.

Author details
[1]Berlin Institute of Health (BIH) at Charité, BIH QUEST Center for Responsible Research, Berlin, Germany. [2]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität Zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany. [3]Department of Clinical Neuroscience, Karolinska Institute, Stockholm, Sweden. [4]Department of Psychology, Stockholm University, Stockholm, Sweden. [5]Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zürich, Switzerland. [6]Medical University of Vienna, CEMSIIS, Section for Medical Statistics, Wien, Austria. [7]Department of Anesthesiology, Intensive Care and Pain Medicine, University Hospital Muenster, Münster, Germany. [8] Animal Welfare Division, University of Bern, Liebefeld, Switzerland. [9]Department of Medical Statistics, University Medical Center, Göttingen, Germany. [10]Institute of Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany. [11]Institute of AI and Informatics in Medicine, Technical University of Munich, TUM School of Medicine, München, Germany. [12]Leibniz Institute on Aging – Fritz-Lipmann Institute, Jena, Germany. [13] Department of Neuroscience and Physiology, Section for Clinical Neuroscience, University of Gothenburg, Gothenburg, Sweden. [14]BIH-visiting Professor at Charité – Universitätsmedizin Berlin, Institute of Public Health, Berlin, Germany. [15]Institute of Clinical and Molecular Virology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany. [16]Institut Für Klinische Epidemiologie Und Angewandte Biometrie (IKEaB) Tübingen Universität, Tübingen, Germany. [17]Biomedical Primate Research Centre (BPRC), Department of Virology, Rijswijk,

The Netherlands. [18]Applied Statistics Group, Department of Mathematics, University of Zurich, Zürich, Switzerland. [19]Division of Clinical Pharmacology, Department of Medicine IV, Klinikum Der Universität München, Munich, Germany. [20]German Cancer Consortium (DKTK), Partner Site, Munich, Germany. [21]Charité -Universitätsmedizin Berlin, FEM & Exzellenzcluster NeuroCure, Animal Behavior Phenotyping Facility, Berlin, Germany. [22]Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Jena, Germany. [23]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig Maximilian University of Munich, Munich, Germany. [24]Institute for Psychopharmacology, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [25]Department of Molecular Neuroimaging, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany. [26]Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, University of Freiburg, Freiburg, Germany. [27]Department of Pediatrics and Children's Cancer Research Center, TUM School of Medicine, Technical University of Munich, Kinderklinik München Schwabing, München, Germany. [28]Kinderklinik München Schwabing- Klinik Und Poliklinik Für Kinder- Und Jugendmedizin, Klinikum Schwabing, München Klinik GmbH und Klinikum Rechts Der Isar (AöR) der Technischen Universität München, München, Germany.

Received: 13 July 2022 Accepted: 21 October 2022
Published online: 07 November 2022

References

1. Bespalov A, Bernard R, Gilis A, Gerlach B, Guillén J, Castagné V, et al. Introduction to the EQIPD quality system. Elife. 2021;10:e63294.
2. Drude NI, Gamboa LM, Danziger M, Dirnagl U, Toelch U. Science Forum: Improving preclinical studies through replications. Elife. 2021;10:e62101.
3. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Elife. 2021;10:e71601.
4. Hunniford VT, Grudniewicz A, Fergusson DA, Grigor E, Lansdell C, Lalu MM. Multicenter preclinical studies as an innovative method to enhance translation: a systematic review of published studies. bioRxiv. 2019:591289.
5. BMBF-DLR. Confirmatory Preclinical Studies (Förderung von konfirmatorischen präklinischen Studien). German Federal Ministry of Education and Research, https://www.gesundheitsforschung-bmbf.de/de/8344.php. German Federal Ministry of Education and Research. 2018. https://www.gesundheitsforschung-bmbf.de/de/8344.php.
6. Krzywinski M, Altman N. Power and sample size. Nat Methods. 2013;10:1139–40.
7. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;2:e124.
8. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science. 2014;1:140216.
9. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12:179–85.
10. Danziger M, Dirnagl U, Toelch U. Increasing discovery rates in preclinical research through optimised statistical decision criteria. bioRxiv. 2022:2022.01.17.476585. https://doi.org/10.1101/2022.01.17.476585.
11. Percie du Sert N, Hurst V, Ahluwalia A, Alam S, Avey MT, et al. The ARRIVE guidelines 2.0: updated guidelines for reporting animal research. PLoS Biol. 2020;18(7):e3000410. https://doi.org/10.1371/journal.pbio.3000410.
12. Kang H. Statistical messages from ARRIVE 2.0 guidelines The Korean. Journal of Pain. 2021;34:1.
13. Festing MFW. The "completely randomised" and the "randomised block" are the only experimental designs suitable for widespread use in preclinical research. Sci Rep. 2020;10:17577.
14. Bespalov A, Wicke K, Castagné V. Blinding and Randomization. In: Bespalov A, Michel MC, Steckler T, editors. Good Research Practice in Non-Clinical Pharmacology and Biomedicine. Cham: Springer International Publishing; 2020. p. 81–100.
15. Parady G, Ory D, Walker J. The overreliance on statistical goodness-of-fit and under-reliance on model validation in discrete choice models: A review of validation practices in the transportation academic literature. Journal of Choice Modelling. 2021;38:100257.
16. Willyard C. Squeaky clean mice could be ruining research. Nature. 2018;556:16–8.
17. Rosshart Stephan P., Herz Jasmin, Vassallo Brian G., Hunter Ashli, Wall Morgan K., Badger Jonathan H., et al. Laboratory mice born to wild mice have natural microbiota and model human immune responses. Science. 2019;365:eaaw4361.
18. Noble H, Heale R. Triangulation in research, with examples. Evid Based Nurs. 2019;22:67.
19. Li C-L, Yu Y, He T, Wang R-R, Geng K-W, Du R, et al. Validating Rat Model of Empathy for Pain: Effects of Pain Expressions in Social Partners. Front Behav Neurosci. 2018;12:242.
20. Smith ML, Hostetler CM, Heinricher MM, Ryabinin AE. Social transfer of pain in mice. Sci Adv. 2016;2:e1600855.
21. Dirnagl U. Preregistration of exploratory research: Learning from the golden age of discovery. PLoS Biol. 2020;18:e3000690.
22. Soderberg CK, Errington TM, Schiavone SR, Bottesini J, Thorn FS, Vazire S, et al. Initial evidence of research quality of registered reports compared with the standard publishing model. Nat Hum Behav. 2021;5:990–7.
23. von Kortzfleisch VT, Karp NA, Palme R, Kaiser S, Sachser N, Richter SH. Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' Sci Rep. 2020;10:16579.
24. Sulaiman A, Wang L. Bridging the divide: preclinical research discrepancies between triple-negative breast cancer cell lines and patient tumors. Oncotarget. 2017;8(68):113269–81. https://doi.org/10.18632/oncotarget.22916. Published 2017 Dec 4.
25. Faria CC, Cascão R, Custódia C, Paisana E, Carvalho T, Pereira P, et al. Patient-derived models of brain metastases recapitulate human disseminated disease. Cell Reports Medicine. 2022;3:100623.
26. Hou X, Du C, Lu L, Yuan S, Zhan M, You P, et al. Opportunities and challenges of patient-derived models in cancer research: patient-derived xenografts, patient-derived organoid and patient-derived cells. World Journal of Surgical Oncology. 2022;20:1–9.
27. Ehlerding EB, Grodzinski P, Cai W, Liu CH. Big Potential from Small Agents: Nanoparticles for Imaging-Based Companion Diagnostics. ACS Nano. 2018;12:2106–21.
28. Valla V, Alzabin S, Koukoura A, Lewis A, Nielsen AA, Vassiliadis E. Companion Diagnostics: State of the Art and New Regulations. Biomark Insights. 2021;16:11772719211047764–11772719211047764.
29. Frommlet F, Heinze G. Experimental replications in animal trials. Lab Anim. 2021;55:65–75.
30. Piper SK, Grittner U, Rex A, Riedel N, Fischer F, Nadon R, et al. Exact replication: Foundation of science or game of chance? PLoS Biol. 2019;17:e3000188.
31. Permanent Senate, Commission on Animal Protection, and Experimentation of the DFG. Animal Experimentation in Research: the 3Rs Principle and the Validity of Scientific Research. 2019. https://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/handreichung_sk_tierversuche_en.pdf.
32. Official Journal of the European Union. DIRECTIVE 2010/63/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 September 2010 on the protection of animals used for scientific purposes (Text with EEA relevance). 2010.
33. Cohen H. The animal welfare act. J animal l. 2006;2:13.
34. Nosek BA, Errington TM. What is replication? PLoS Biol. 2020;18:e3000691.
35. Kimmelman J, Mogil JS, Dirnagl U. Distinguishing between Exploratory and Confirmatory Preclinical Research Will Improve Translation. PLoS Biol. 2014;12:e1001863.
36. Voelkl B, Altman NS, Forsman A, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020;21:384–93. https://doi.org/10.1038/s41583-020-0313-3.
37. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol. 2018;16:e2003693.
38. von Kortzfleisch VT, Ambrée O, Karp NA, Meyer N, Novak J, Palme R, et al. Do multiple experimenters improve the reproducibility of animal studies? PLoS Biol. 2022;20:e3001564.
39. Nigri M, Åhlgren J, Wolfer DP, Voikar V. Role of environment and experimenter in reproducibility of behavioral studies with laboratory

mice. Front Behav Neurosci. 2022;16. https://www.frontiersin.org/articles/10.3389/fnbeh.2022.835444.
40. Percie du Sert N, Rice ASC. Improving the translation of analgesic drugs to the clinic: animal models of neuropathic pain. British Journal of Pharmacology. 2014;171:2951–63.
41. Jaggi AS, Jain V, Singh N. Animal models of neuropathic pain. Fundam Clin Pharmacol. 2011;25:1–28.
42. Munafò MR, Smith GD. Robust research needs many lines of evidence. Nature. 2018;553:399–401.
43. Metselaar JM, Lammers T. Challenges in nanomedicine clinical translation. Drug Deliv Transl Res. 2020;10:721–5.
44. Balasubramanian B, Venkatraman S, Myint KZ, Janvilisri T, Wongprasert K, Kumkate S, et al. Co-clinical trials: an innovative drug development platform for cholangiocarcinoma. Pharmaceuticals. 2021;14:51. https://doi.org/10.3390/ph14010051.
45. Bonapersona V, Hoijtink H, Abbinck M, Baram TZ, Bolton JL, Bordes J, et al. Increasing the statistical power of animal experiments with historical control data. Nat Neurosci. 2021;24:470–7.
46 Bonapersona V, Hoijtink H, Joëls M. Sarabdjitsingh RA. P.201 Reduction by Prior Animal Informed Research (RePAIR): a power solution to animal experimentation. European Neuropsychopharmacology. 2020;31:S19-20.
47. Wassmer G. On sample size determination in multi-armed confirmatory adaptive designs. null. 2011;21:802–17.
48. Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. Stat Med. 2012;31:4269–79.
49. Kramer M, Font E. Reducing sample size in experiments with animals: historical controls and related strategies. Biol Rev. 2017;92:431–45.
50. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 2007;82:591–605.
51. Bakker A, Cai J, English L, Kaiser G, Mesa V, Van Dooren W. Beyond small, medium, or large: points of consideration when interpreting effect sizes. Educ Stud Math. 2019;102:1–8.
52. Rosnow RL, Rosenthal R. Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychol Methods. 1996;1:331.
53. Hirst JA, Howick J, Aronson JK, Roberts N, Perera R, Koshiaris C, et al. The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. PLoS ONE. 2014;9:e98856.
54. Leenaars CHC, Kouwenaar C, Stafleu FR, Bleich A, Ritskes-Hoitinga M, De Vries RBM, et al. Animal to human translation: a systematic scoping review of reported concordance rates. J Transl Med. 2019;17:223.
55. Held L. The assessment of intrinsic credibility and a new argument for p < 0.005. R Soc Open Sci. 2019;6:181534–181534.
56. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nature Human. Behaviour. 2018;2:6–10.
57. Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, et al. Justify your alpha. Nature Human. Behaviour. 2018;2:168–71.
58. Peder M. Isager, Robbie C. M. van Aert, Bahník Š, Mark J. Brandt, Kurt A. DeSoto, Roger Giner-Sorolla, et al. Deciding what to replicate: A decision model for replication studyselection under resource and knowledge constraints. MetaArXiv. 2020. https://doi.org/10.31222/osf.io/2gurz.
59. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or predictive power? Control Clin Trials. 1986;7:8–17.
60. Neumann K, Grittner U, Piper SK, Rex A, Florez-Vargas O, Karystianis G, et al. Increasing efficiency of preclinical research by group sequential designs. PLoS Biol. 2017;15:e2001307.
61. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci. 2014;9:641–51.
62. Senn SS. Statistical issues in drug development. 3rd ed: Wiley; 2008. ISBN: 978-1-119-23857-7.
63. Amaral OB, Neves K, Wasilewska-Sampaio AP, Carneiro CF. The Brazilian Reproducibility Initiative eLife. 2019;8:e41602.
64. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. Elife. 2014;3:e04333.
65. Nosek BA, Errington TM. Making sense of replications. Elife. 2017;6:e23383.
66. Held L. A new standard for the analysis and design of replication studies. J R Stat Soc A Stat Soc. 2020;183:431–48.
67. Vogel AL, Knebel AR, Faupel-Badger JM, Portilla LM, Simeonov A. A systems approach to enable effective team science from the internal research program of the National Center for Advancing Translational Sciences. Journal of Clinical and Translational Science. 2021;5:e163.

Publisher's Note