



Association of Animal Models in the Field of Translational Medicine: Prediction and Validation

46

Debora Bizzaro

Contents

Introduction	894
Validity of Preclinical Animal Research	895
Internal Validity	895
External Validity	895
Association of Internal and External Validity	896
How to Refine the Preclinical Animal Models	897
Conclusion	899
References	899

Abstract

During clinical development, the success rates for drugs remain low despite large investments. A prominent reason for the poor rate of translation from bench to bedside is generally assumed to be the failure of preclinical animal models to predict clinical efficacy and safety. This failure could be held to problems of internal validity (e.g., poor study design, lack of measures to control biases) and external validity (e.g., poor reproducibility of a research finding, translational failure) in preclinical animal studies. To analyze the significance of animal research impartially, we must warrant that (1) the experiments are conducted and reported according to best scientific practices; (2) the selection of animal models is made with a clear and thorough translational rationale behind it. Once these criteria are met, the true significance of the use of animal models in drug development can be ultimately attested.

D. Bizzaro

Multivisceral Transplant Unit-Gastroenterology, Department of Surgery, Oncology and Gastroenterology, University Hospital of Padova-Italy, Padova, Italy
e-mail: debora.bizzaro@gmail.com; debora.bizzaro@unipd.it

Keywords

Preclinical animal models · Internal validity · External validity · Experimental bias

Introduction

During clinical development, the success rates for drugs remain low despite large investments. During the last decade, thanks to innovative molecular biology approaches, a lot of possible new drug targets were identified and a large number of promising drug candidates directing these potential new targets have entered clinical development. However, this was associated to the decline in the success rates for drugs due to the absence of efficacy in humans during clinical trials (De Martini 2020; Hwang et al. 2016). On average, 17 years pass before effective research-based findings are applied into practice, and even then only an estimated 14% of those findings result in health care delivery changes (Olson and Oudshoorn 2020).

A complex multifactorial process is at the base of the poor rates of successful translation from bench to bedside and includes, as one of the most important components, the failure of preclinical animal models in predicting clinical efficacy and safety (Ferreira et al. 2020).

More strict success criteria could be useful to decrease the failure rates during the preclinical studies, especially during target validation for a potential new drug. Efforts to explain these failures have focused on the internal and external validity of preclinical animal models (Henderson et al. 2013; Ferreira et al. 2020).

So far, the internal validation is the most considered and analyzed aspect of the problem. It included inadequacy in study design such as low power, inappropriate endpoints and inaccuracies in study conduction, analysis, and reporting. These inaccuracies lead to unreliable data which ultimately means unnecessary suffering for the animals and potential risks for human clinical trial participants (Ioannidis 2017).

On the contrary, there has been relatively little discussion of the other key factor influencing translation and assessing the reliability of the research, namely the external validity. It is defined as the degree to which research results derived in one experiment setting, species, or strains can be consistently applied to other different settings, species, or strains (Pound and Ritskes-Hoitinga 2018). Of course, in the field of preclinical animal research, external validity is of the utmost importance.

When adequately designed and conducted, preclinical animal studies with adequate internal and external validation are essential in the discovery and development of new drugs. The current chapter highlights some strategies that researchers should take into account to improve the translational potential of the preclinical animal models.

Validity of Preclinical Animal Research

Internal Validity

Internal validity refers to the basic principles of study design, such as conduction, analysis, and results reporting. Frequently, preclinical animal studies suffer from serious problems of internal validity, in particular low power, inappropriate endpoints, and lack of measures to avoid bias such as randomization or blinding (Denayer et al. 2014).

Recently, Jankovic et al. found that only 7% of the published studies they analyzed from journals with relatively high impact factors retained strong internal validity, despite all having undergone a rigorous review process. They highlighted two major flaws in internal validity: lack of randomization and the use of pseudo-replication (repeating an experiment using the same animal) that can lead to misleading data and overestimation of the sample size (Jankovic et al. 2019). Similarly, other research found that only 14% of the papers reported blinding in animal selection and results in evaluation to avoid bias (Kilkenny et al. 2009). Moreover, only 3% of all studies reported sample size and statistical power calculation and more less studies defined a primary outcome variable (Macleod 2011).

Therefore, to avoid biased conclusions and production of false positive or negative results, the preclinical studies with animal models should, at least, contain the appropriate control groups, be repeated in independent sets of animals with adequate statistical power to guarantee significant results, and apply treatment randomization, as well as a blinded outcome assessment (Pound and Ritskes-Hoitinga 2018; Schmidt-Pogoda et al. 2020).

External Validity

The external validity extends behind the specific setting of experiment and is referred to as the grade of generalization of study results, i.e., how replicable they are in other environmental conditions, experimental setting, study populations, and even in other strains or species of animals, including humans (Pound and Ritskes-Hoitinga 2018). Poor external validity may implicate: i) poor reproducibility of a research finding (e.g., the same experiment in a different laboratory by a different investigator produces different results); ii) translational failure (e.g., an effective treatment in an animal model has not therapeutic effect in a human clinical trial).

McKinney and Bunney were the first to recommend criteria on external validity of animal models in 1969, mainly focusing on affective disorders (McKinney and Bunney 1969). In 1984, these external validations were simplified to three criteria: predictive, face, and construct validity (Willner 1984). These are the most widely accepted criteria for model validation, although others were proposed (Denayer et al. 2014; Ferreira et al. 2020).

Predictive validity is defined as the determination of how preclinical animal models are effective in predicting currently unknown aspects of the human disease or the clinical efficacy of a drug.

Face validity refers to the likeness in pathobiology, symptoms, and signs among animal models and human diseases. In many cases, the pathobiology underlying the symptoms of the disease is poorly understood, so the assessment of facial validity is often hampered.

Construct validity is defined as how the method of induction of the disease phenotype in animals replicates the currently known disease etiology in humans (similarity in the biological dysfunction).

By definition, a model cannot be a perfect reproduction of the human disease. Consequently, all the three criteria cannot be met by only one model; for example, a model might have sound predictive validity but totally lack face validity, or vice versa. A combination of diverse animal models could be surely more similar to the clinical situation than a single complex model.

The three criteria offer a general external validation and there is controversy in their ranked importance, mostly due to the discrepancy in their definition. Generally, their importance should be based on the purpose of the model. Indeed, according to the aim of the animal model, the criteria to be respected may change. For example, face validity may be more important in animal models for pathobiology studies, whereas in preclinical drug discovery, predictive validity tends to hold the most weight. Understanding which validity a model can and cannot provide is fundamental for accurate preclinical assessment of novel therapeutic agents (Denayer et al. 2014).

Association of Internal and External Validity

Despite it is a common perception that internal validity is actually a prerequisite of external validity (i.e., by resolving the problems of internal validity the clinical translation would be more successful), the available evidence does not support this sight. Indeed, it is important to highlight that preclinical animal studies need to be both internally and externally validated if they have to be translated into benefits for humans (van der Worp et al. 2010). Both internal and external validity are critical, yet researchers often encounter a trade-off between them, such that strengthening the features of one type of validity weakens the other. Some of the strategies used to increase internal validity could together decrease external validity. For example, using homogeneous study populations to standardize experiments and to maximize test sensitivity inexorably prejudice the external validity of the findings, resulting in poor reproducibility (van der Worp et al. 2010). Preclinical studies are usually performed in a fairly homogeneous approach (e.g., mice of the same sex, age, and genetic background). Despite this may ease the use of as a small number of animals as possible to obtain a statistically significant result, it does not really represent the real human condition of a pool of individuals from various genetic and environmental backgrounds. Results more applicable in spite of the animal's (or human's)

characteristics would be reached by combining a heterogeneous population of subjects with the right analytical techniques (Pound and Ritskes-Hoitinga 2018).

How to Refine the Preclinical Animal Models

Taking into account both internal and external validity, numerous aspects should be weighed in performing preclinical studies.

- I. Selection and attrition bias: refer to the biased distribution of animals to treatment and control groups and can be prevented by randomization. Randomization is always required even if homogeneous population (such as same sex and/or age, inbred mice kept under identical housing conditions) was used, since individual differences still prevail. Since selection bias may occur either consciously or subconsciously, operator-independent methods may be preferable (e.g., random number generators). Selection biases may also occur if animals' inclusion or exclusion criteria are weakly defined. Complications that require exclusion of animals (e.g., reaching of humane endpoints or occurring of complications unrelated to the experimental treatment that make the outcome analysis worthless) are an intrinsic risk in preclinical animal studies. To avoid this kind of bias, all animal inclusion and exclusion criteria should be clearly predefined, and the operator accountable of these steps should be unaware of the treatment allocation (van der Worp et al. 2010). The risk, if these criteria are not well specified, is the unequal distribution of withdraws among treatment groups, defined as attrition bias.
- II. Performance and detection biases: the first occurs when there is a systematic difference in the animal care and/or experimental procedures (apart from the treatment under investigation) between the treatment groups. Detection bias occurs when the outcomes are determined differently in animals of distinct treatment groups. Both these biases may occur consciously or subconsciously, therefore the best approach to exclude them is blinding. In contrast to randomization, blinding is not always achievable and it is essential that authors explicitly report the blinding status of the staff involved in experimental steps that may affect the outcome of the study (Denayer et al. 2014).
- III. Sample size and power analysis: according to one of the 3R principles (reduction) the researcher should minimize the number of animals utilized in biomedical experimentations. However, this should be well-adjusted with the statistical power essential to obtain relevant data (Button et al. 2013). When possible, sample size calculation and power analysis should be carried out, specifying the desired statistical power, the level of statistical significance, and the minimal effect size considered to be relevant.
- IV. Reproducibility of results: generally, experimental set-ups could be highly standardized in a single lab, while minimal differences in environment (such as staff, noise) or experimental procedures (e.g., a xenograft model with a

- diverse cell line) in another laboratory may harvest important differences in results avoiding their generalization potential in a wider context (Richter 2017).
- V. Treatment time course: treatment of animal models is often started very shortly after or even before the disease onset. In this condition, the treatment is prophylactic, evidently in contrast to the human “real-life” condition in which the treatment is usually therapeutic, therefore started only after the clear manifestation of the symptoms and diagnosis. As a consequence, in an animal model the potential pharmacological effect may be wrongly overestimated (Malfait and Little 2015).
 - VI. Animal species and strain: The selection of species and strain of animals for a particular model should be carefully performed. Primarily, the animal target should be sensible to the active principle of the drug to be tested. In addition, the health status, age, and gender of the animals should be matched as strongly as possible to the “real-life” clinical condition. Instead, the animals used in preclinical research be likely more young and healthy, while numerous human diseases develop in older age and in association with other co-morbidities (Malfait and Little 2015). Furthermore, many animal models do not have the complexity necessary to precisely reproduce human conditions. Clearly, in such cases the findings from animal studies can give misleading results and are improbable to be appropriate for human patients.
 - VII. Reporting: together with the problem of insufficient reporting of experimental procedures that limits the reproducibility of the same experiment, a further obstacle is that experiments with positive and statistically significant results are more likely to be disseminated to the scientific community than negative ones. This is due to selective analysis and selective outcome reporting. Selective analysis occurs when numerous statistical analyses are performed but the Authors present only the one with the most statistically significant result; selective outcome reporting occurs when numerous result variables are analyzed but only the ones that are significantly influenced by the treatment are reported (Tsilidis et al. 2013). To avoid these potential biases, primary and secondary outcome variables as well as the statistical approaches to testing for treatment effects should be defined before the onset of the study.
 - VIII. Efficacy and safety assessment: efficacy is generally analyzed in preclinical disease models treated with a therapeutic dose of the drug but without examination of side effects, while safety is assessed in healthy animals to whom was administered the drug at high dose. A safety margin is then defined by comparing the effective doses to that outlined in the safety assessment. However, this margin might be overestimated. Indeed, healthy and young animals used for safety analysis could develop less potential side effects as compared to diseased and more frail subjects. On the other hand, estimating the efficacy without considering the side effect could make it impossible to administer corresponding doses in a clinical setting. A possible solution would be to use diseased animals in parallel to standard healthy animals for safety testing of the new drugs.

Conclusion

Animal models are an essential aspect of any drug development experiment. However, inaccuracies in experimental design, conduction, and publication (whether conscious or not) persist to afflict research based on animal models. Facing these problems and underlying causes is an essential step in the direction of successful improvement of experimental design and conduct. Researchers, but also reviewers, and journal editors should not only support such methods of refinement but rigorously implement them. Otherwise, the reliability and ethical justification of animal research may be permanently damaged.

References

- Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*:365–376. <https://doi.org/10.1038/nrn3475>. England
- De Martini D (2020) Empowering phase II clinical trials to reduce phase III failures. *Pharm Stat* 19 (3):178–186. <https://doi.org/10.1002/pst.1980>
- Denayer T, Stöhr T, Van Roy M (2014) Animal models in translational medicine: validation and prediction. *New Horizons Transl Med* 2(1):5–11. <https://doi.org/10.1016/j.nhtm.2014.08.001>
- Ferreira GS, Veening-Griffioen DH, Boon WPC, Moors EHM, van Meer PJK (2020) Levelling the translational gap for animal to human efficacy data. *Animals (Basel)* 10(7). <https://doi.org/10.3390/ani10071199>
- Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG (2013) 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(7):e1001489. <https://doi.org/10.1371/journal.pmed.1001489>
- Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS (2016) Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern Med* 176(12):1826–1833. <https://doi.org/10.1001/jamainternmed.2016.6008>
- Ioannidis JP (2017) Acknowledging and overcoming nonreproducibility in basic and preclinical research. *JAMA* 317(10):1019–1020. <https://doi.org/10.1001/jama.2017.0549>
- Jankovic SM, Kapo B, Sukalo A, Masic I (2019) Evaluation of published preclinical experimental studies in medicine: methodology issues. *Med Arch* 73(5):298–302. <https://doi.org/10.5455/medarh.2019.73.298-302>
- Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D et al (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One* 4(11):e7824. <https://doi.org/10.1371/journal.pone.0007824>
- Macleod M (2011) Why animal research needs to improve. *Nature*:511. <https://doi.org/10.1038/477511a>. England
- Malfait AM, Little CB (2015) On the predictive utility of animal models of osteoarthritis. *Arthritis Res Ther* 17(1):225. <https://doi.org/10.1186/s13075-015-0747-6>
- McKinney WT Jr, Bunney WE Jr (1969) Animal model of depression. I. Review of evidence: implications for research. *Arch Gen Psychiatry* 21(2):240–248. <https://doi.org/10.1001/archpsyc.1969.01740200112015>
- Olson A, Oudshoorn A (2020) Knowledge translation: a concept analysis. *Nurs Forum* 55(2):157–164. <https://doi.org/10.1111/nuf.12410>
- Pound P, Ritskes-Hoitinga M (2018) Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. *J Transl Med* 16(1): 304. <https://doi.org/10.1186/s12967-018-1678-1>

- Richter SH (2017) Systematic heterogenization for better reproducibility in animal experimentation. *Lab Anim (NY)* 46(9):343–349. <https://doi.org/10.1038/labani.1330>
- Schmidt-Pogoda A, Bonberg N, Koecke MHM, Strecker JK, Wellmann J, Bruckmann NM et al (2020) Why most acute stroke studies are positive in animals but not in patients: a systematic comparison of preclinical, early phase, and phase 3 clinical trials of neuroprotective agents. *Ann Neurol* 87(1):40–51. <https://doi.org/10.1002/ana.25643>
- Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW et al (2013) Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biol* 11(7):e1001609. <https://doi.org/10.1371/journal.pbio.1001609>
- van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V et al (2010) Can animal models of disease reliably inform human studies? *PLoS Med* 7(3):e1000245. <https://doi.org/10.1371/journal.pmed.1000245>
- Willner P (1984) The validity of animal models of depression. *Psychopharmacology* 83(1):1–16. <https://doi.org/10.1007/bf00427414>