Do animal study metrics predict human trial success?

Written by The Life Science Feed Editorial Team · Published 19 August 2026 · Edited by Mara Voss

> Translating preclinical animal research into successful human clinical trials, where the animal findings lead to effective human treatments, remains a persistent challenge in biomedical science. The frequent failure of animal findings to reproduce in humans wastes resources and delays therapeutic advancements. This phenomenon, termed translation failure, highlights a fundamental disconnect in how we evaluate early-stage evidence.1
> A simulation study published in eLife examined whether established metrics for replication success, typically applied to human-to-human study comparisons, can accurately quantify animal-to-human translation.1 The findings suggest that while some metrics offer utility, their performance varies significantly with underlying evidence and assumptions, often failing to control false positive rates under conditions of heterogeneity.1

KEY TAKEAWAYS

- The Pivot: Most replication success metrics, except meta-analysis and replication Bayes factor, failed to control false positive rates when heterogeneity between human studies increased.
- The Data: Translation power, the probability of true positive translation success, was constrained by the weaker evidence, for example, small animal sample sizes resulted in lower power.1
- The Action: Clinicians should view animal-to-human translation claims with caution, especially when studies lack robust evidence or exhibit significant heterogeneity; relying on multiple metrics is advisable.

The biomedical research community has long grappled with the issue of translation failure, where seemingly robust findings from animal models do not hold up in human trials. This problem is not merely an academic concern; it directly impacts the pipeline of new therapies reaching patients and the efficient allocation of research funding. Understanding the limitations of current evaluative tools is therefore critical for improving the predictability of preclinical research.1 Huang, Pawel, and Wever conducted a comprehensive simulation study to evaluate the applicability of nine distinct replication success metrics in the context of animal-to-human translation.1 The researchers aimed to determine if these metrics, designed for assessing reproducibility between human studies, could effectively quantify the success or failure of translating animal findings to human outcomes. They simulated animal and human studies under 648 varied scenarios, drawing parameters from a meta-analysis on prenatal amino acid supplementation and maternal blood pressure.1 The scenarios explored different effect sizes, levels of heterogeneity, animal sample sizes, and the number of pooled animal studies.1

The Metrics Under Scrutiny

The nine metrics assessed included the two-trials rule, meta-analysis, replication Bayes factor, unweighted and weighted Edgington's methods, the golden skeptical p-value, and three versions of the controlled skeptical p-value.1 Each of these metrics approaches the concept of replication or translation success from a slightly different statistical angle,

offering a range of perspectives on how to interpret agreement between studies. The two-trials rule, for instance, is a straightforward approach requiring both the original and replication studies to show statistical significance in the same direction. Meta-analysis, conversely, pools evidence to provide a combined effect estimate, often considered a gold standard for synthesizing evidence.1

Replication Bayes factors quantify the evidence for the replication hypothesis over the null hypothesis, offering a continuous measure of support. Edgington's methods, both unweighted and weighted, combine p-values from multiple studies. Skeptical p-values, including the golden and controlled versions, are designed to be more conservative, requiring stronger evidence from the replication study to declare success, especially when the original finding was barely significant. These varied approaches highlight the complexity in defining and measuring 'success' in a translational context, a challenge that has been explored in other areas of clinical trial design, such as how retracted studies impact efficacy assessments.1

Performance Under Varying Conditions

The simulation revealed critical insights into the performance of these metrics. Most metrics, with the notable exceptions of meta-analysis and the replication Bayes factor, controlled false positive rates effectively when there was no heterogeneity.1 This means they correctly identified true negatives, avoiding declarations of translation success when none existed. But this control deteriorated significantly as heterogeneity increased, particularly between human studies.1 In such scenarios, these metrics became liberal, frequently indicating success when it was not warranted. This finding is particularly concerning given the inherent variability often observed between animal models and human physiology, as well as between different human populations or study designs.1

Translation power, defined as the probability of achieving true positive translation success, consistently suffered when the evidence from either the animal or human studies was weak.1 For example, small sample sizes in the animal studies directly resulted in lower translation power, regardless of the human study's strength. This highlights a fundamental principle: the chain of evidence is only as strong as its weakest link. If preclinical studies are underpowered or poorly designed, even impeccably conducted human trials will struggle to demonstrate translation success, a point relevant to discussions around novel targets in RNA metabolism where early data can be highly variable.1

The metric based on meta-analysis frequently indicated success when either of the species found strong evidence.1 This approach, by pooling data, can sometimes mask discrepancies or overstate the consistency of findings, especially if one study is overwhelmingly large or has a very precise estimate. Skeptical p-values, on the other hand, proved more conservative. They demanded a higher bar for declaring translation success, which, while potentially reducing false positives, might also increase false negatives, missing genuine translational signals that are less overtly strong.1

Identifying More Consistent Performers

Among the metrics evaluated, the skeptical p-value that controls overall type-one error and the weighted version of Edgington's method performed relatively consistently across the diverse scenarios.1 These metrics demonstrated a better balance between controlling false positives and maintaining reasonable translation power, even in the presence of heterogeneity. Still, no single metric emerged as uniformly optimal. Each had its strengths and limitations, performing better under specific conditions of effect size, heterogeneity, and sample size. This lack of a 'silver bullet' metric suggests that the complexity of animal-to-human translation cannot be captured by a single statistical test.1

The study's reliance on parameters from a meta-analysis on prenatal amino acid supplementation and maternal blood pressure provides a clinically relevant context for the simulations.1 This specific area, involving physiological outcomes and interventions, offers a realistic backdrop for evaluating how well these metrics perform. But the generalizability of these findings to other therapeutic areas, particularly those with different biological complexities or higher inherent variability, remains an open question. For instance, the challenges in defining success metrics are also evident in areas like LVAD outcomes, where survival is clear but quality of life metrics are harder to capture.1

The Catch: Limitations and Future Directions

The simulation study, while comprehensive, is inherently limited by its reliance on simulated data. While the parameters were drawn from real-world meta-analyses, simulations cannot fully capture the myriad unmeasured confounders and biological intricacies that characterize actual animal and human studies. The choice of 648 scenarios, though extensive, represents a finite subset of all possible conditions. The performance of these metrics might differ under extreme or unusual circumstances not covered in the simulations.1

The authors recommend using multiple metrics in combination, paying close attention to their individual strengths and limitations, when evaluating the translation of animal findings to human outcomes.1 This multi-faceted approach acknowledges that no single statistical tool can perfectly encapsulate the relationship between preclinical and clinical evidence. For clinicians, this means maintaining a healthy skepticism towards claims of 'replication' or 'translation success' based on a single metric, especially when the underlying evidence base is small or highly variable. The Oxford Handbook of General Practice often emphasizes the need for a critical appraisal of evidence, a principle that applies equally to preclinical data.1

The study also highlights the ongoing need for more rigorous preclinical research designs, particularly concerning sample size determination and the reporting of heterogeneity. If animal studies are consistently underpowered, or if their heterogeneity is not adequately accounted for, even the most sophisticated replication metrics will struggle to provide reliable assessments of translational potential. This issue is not unique to animal studies; similar challenges arise in human trials, where assuming success based on AFib type alone can lead to misinterpretations.1

> **CLINICAL IMPLICATIONS**
>
> The persistent challenge of translating animal research into human clinical success demands a more critical eye from clinicians. When a new therapy emerges from preclinical studies, the enthusiasm must be tempered by a clear understanding of how 'success' was measured in the animal-to-human bridge. This simulation study confirms that many standard metrics are simply not up to the task, particularly when biological variability, or heterogeneity, is high.
>
> For general practitioners and specialists alike, this means that a positive animal study result, even if statistically 'replicated' by some metrics, does not automatically confer a high probability of human benefit. The inherent differences between species, coupled with the often-liberal nature of certain metrics under heterogeneity, can lead to false positives that waste research funding and raise false hopes. We must demand more robust evidence and a transparent discussion of the metrics used to declare translational success.
>
> Industry sponsors and academic researchers should adopt a more conservative, multi-metric approach to evaluating preclinical data. Relying on a single, potentially flawed metric risks pushing therapies into human

> trials that are unlikely to succeed, diverting resources from more avenues that show greater likelihood of success. The skeptical p-value that controls overall type-one error and the weighted Edgington's method offer more consistent performance, but even these are not perfect. A combination of approaches, with a clear understanding of their limitations, is essential.

EDITORIAL & AI STANDARDS

AI tools are used solely to structure and summarise evidence from peer-reviewed primary sources. No AI-generated conclusions appear in The Life Science Feed without editor verification against the primary source.

REFERENCES

1. 1. Huang CJ, Pawel S, Wever KE. Evaluating the applicability of replication success metrics in animal-to-human translation: A simulation study. Elife. 2026;15:e42565459. doi:10.7554/eLife.42565459

LICENCE & RIGHTS

MEDICAL DISCLAIMER

This content is intended for healthcare professionals only and does not constitute medical advice. Clinical decisions must be made by qualified practitioners based on individual patient circumstances.

FOLLOW THE LIFE SCIENCE FEED

Instagram @lifesciencefeed **LinkedIn** linkedin.com/company/thelifesciencefeed

thelifesciencefeed.com