Consensus paper II: Role of Data Reproducibility in Biomedical Technology Translation

Draft Perspective for Science Translational Medicine on Translate! 2021 Discussion

Ulrich Dirnagl[1,2*], Georg N Duda[3*], David W. Grainger[4,5*], Petra Reinke[6*], Ronenn Roubenoff[7*]

[1] Department of Experimental Neurology, Charité - Universitätsmedizin Berlin, Germany
[2] QUEST Center for Transforming Biomedical Research, Berlin Institute of Health, Germany
[3] Berlin Institute of Health (BIH) Center for Regenerative Therapies (BCRT), BIH-Center Julius Wolff Institute for Biomechanics and Musculoskeletal Regeneration, Institute Medical Immunology, Charité Universitätsmedizin Berlin, 13353, Berlin, Germany
[4] Department of Pharmaceutics and Pharmaceutical Chemistry, Health Sciences, University of Utah, Salt Lake City, UT 84112, USA
[5] Department of Biomedical Engineering, University of Utah, Salt Lake City, UT 84112, USA
[6] Berlin Center for Advanced Therapies (BeCAT), Charité - Universitaetsmedizin Berlin, 13353 Berlin, Germany
[7] Novartis Institutes for Biomedical Research, Cambridge, Massachusetts, and Basel, Switzerland

*names listed in alphabetical order; all authors equally contributed to the paper as first/senior authors

Many challenges are identified in transferring observations from model experimental research systems (e.g., *in silico, in vitro, ex vivo,* or *in vivo* in animals) towards treatments of human diseases. A sound understanding of the medical need and corresponding relevant preclinical data is necessary to confidently proceed from preclinical to clinical testing for validation. The ability to obtain support for essential funding needed to develop candidate therapeutics (small molecules, biomolecules, cell and gene therapies, etc.) or medical technologies to be brought into first-in-human trials and later phases of registration relies on these risky translational research correlations. However, apparently exciting preclinical data may fail to translate to the next step at many levels. Science has had to learn the hard way that, for example, despite voluminous data showing that p38 inhibition successfully treats inflammation in animals, over twenty p38 inhibitors have failed in human trials (Hammaker D, Firestein GS "Go upstream, young man": lessons learned from the p38 saga. Ann Rheum Dis 2010;69:i77-i82 http://dx.doi.org/10.1136/ard.2009.119479). For vascular graft prosthetic development, preclinical calf and non-human primates exhibit greater similarity to humans in their coagulation and fibrinolytic responses than those of dogs or pigs. Yet no model except non-human primates demonstrates the incomplete re-endothelialization that characterizes (and plagues) human vascular implants. Despite acknowledged limitations in human performance comparisons, the dog model was used for over two decades per expert panel recommendation, and only recently rescinded as inappropriate from several studies for assessing many human-relevant implant performance parameters (P. Zilla, D. Bezuidenhout, P., Human prosthetic vascular grafts: wrong models, wrong questions and no healing, Biomaterials, 28 (2007), pp. 5009-5027; **DOI:** 10.1016/j.biomaterials.2007.07.017). To date, the synthetic small diameter vascular graft remains the Holy Grail of vascular implants. Myriad examples of poor preclinical and *in vitro* model relevance to the human condition are published annually. The case for poor predictability of cancer models in human clinical trials is often the poster child for miserable preclinical models (doi: 10.1038/nrd1635). The false hope, compromised research reputation, and enormous wasted expenses both on preclinical

claims for treating myriad tumors, and subsequent clinical failures continues to haunt patients and physicians, while feeding the insatiable appetite for further, perhaps irrelevant, cancer research. Similar things are claimed about preclinical pain models (https://doi.org/10.1016/j.jneumeth.2020.108997; https://doi.org/10.1038/nrn2606). Nonetheless, research continues to move forward, pretending perhaps that none of this really matters. The exact financial and the less exact opportunity costs of continuing with poor preclinical biomedical research models is staggering (PLOS Biology, DOI:10.1371/journal.pbio.1002165; 2015).

Some experimental uncertainties in models are certainly outside the researchers' control – biology is indeed complex, often non-linear, difficult to model and perhaps at times, chaotic. The biology may simply differ in humans compared to model systems or non-human animals. In these cases, research model pursuit can be justified in terms other than direct one-to-one human relevance, such as isolating a specific relevant mechanistic signal, elucidating a single relevant pathway, or assessing a genetic contribution. Some argue that external validity (how reliably research outcomes from a particular experimental design, sample population or selected species, and from a specific laboratory environment translate to other laboratories running similar experiments, in other study populations and also extended to other species) as well as internal validity (competent experimental design, expert conduct and analysis, and accurate reporting) are both essential. A common assertion is that reliable translation of research results from animal models to humans can only occur if preclinical animal studies are both internally and externally valid (https://doi.org/10.1186/s12967-018-1678-1). However, there are additional factors that we in the biomedical research community can better control and incorporate. ***In this brief paper, we suggest some issues that could be better addressed in biomedical research translational pathways via best practices.***

First, **reproducibility and robustness** are the bedrock of research and as such, also of biomedical translation. Meta-research of the past decade has provided overwhelming evidence that research of low internal and statistical power is a major cause of translational attrition. For example, a recent analysis of 1.6 million papers (1997–2019) quantifying the rigor and transparency in the reporting of preclinical research demonstrated that less than 30% of studies mention methods to reduce bias (blinding, randomization, etc.) (https://doi.org/10.1016/j.isci.2020.101698). Average statistical power in most work appears to be below 10%. As a consequence, false positive, as well as false negative results abound, and where effects are real, their effect sizes will be substantially overestimated (https://doi.org/10.1371/journal.pmed.0020124; https://doi.org/10.1098/rsos.140216). Predominant reliance of the biomedical field on null hypothesis significance testing (NHST), and associated use and misuse of *P* values for validation is argued to now be the "most widely perpetrated misdeed of statistical inference across all of science" (Chavalarias, D., Wallach, J. D., Li, A. H., Ioannidis, J. P. (2016), "Evolution of Reporting P Values in the Biomedical Literature, 1990–2015," *JAMA*, 315, 1141–1148.). Selection bias, p-hacking, and data cherry picking are common modalities used to falsely assert statistical validity for study conclusions. Proper application of effect sizes, confidence intervals, techniques analyzing false discovery rates, Bayesian methods, and adoption of more stringent thresholds for asserting *P* values are all proposed alternatives to avoiding these increasingly reported questionable practices that plague experimental reproducibility and data robustness.(https://doi.org/10.1080/00031305.2018.1447512;https://jamanetwork.com/journals/jama/article-abstract/2676503) To make matters worse, strong publication bias is increasingly evident

(https://doi.org/10.1371/journal.pbio.1001609), meaning that a large portion of the experimental evidence not supporting the investigated hypotheses or treatments is missing from the literature.

Second, **generalizability** is relevant for translation success. Here one must consider how good the chosen model reflects factors that are highly relevant for disease pathophysiology, and hence for possible therapy, such as sex, age, but also status of the immune system, microbiome, etc (e.g. doi: 10.3389/fimmu.2019.00797). In-depth knowledge of the patient-specific medical needs, its characteristics and variability is essential. Recently it was proposed that, quite counterintuitively, experimental heterogeneity should be more widely embraced to improve reproducibility and translatability. Instead of increasing reproducibility, the current emphasis on experimental standardization may actually reduce variability within studies and lead to idiosyncratic, lab-specific results that are not reproducible (doi.org/10.1371/journal.pbio.2003693). Deliberately introducing heterogeneity ("heterogenization") into the experimental design may lead to higher success in drug discovery and translation.

The translatability of preclinical models in predicting human results varies dramatically by disease. In many diseases, the primary animal model has been "cured" many times without leading to a successful human therapy or mitigation (e.g., the mdx mouse of Duchenne muscular dystrophy, the EAE-model of Multiple Sclerosis, different animal models for sepsis, different animal models for tolerance induction in Solid Transplantation, hundreds of diverse refractory murine tumor types, many animal models of human bacterial infections, also with implant infections), while other unmet needs manifest better predictivity (adjuvant arthritis models predicting efficacy of TNF inhibitors for rheumatoid arthritis, osteoporosis, immunotherapies involving checkpoint inhibitors and CAR-T cells in liquid tumors), even though the animal model might share few features closely associated with the human disease (e.g., hardly any mammals progress naturally to osteoporosis or spontaneously develop tumors as in humans). Additional data to test a hypothesis before proceeding to human trials ideally should be orthogonally designed – preferably from genetic validation in humans, or from inferential human evidence such as "real world" data, related known pathway interventions by other human drugs, etc. *We contend that spending time, effort and money to create models seeking closer relationships to human disease is not an efficient use of research resources*. Afterall, most animal models have recognized limitations that may never duplicate any human disease entirely. For the biopharmaceutical industry, it is much more effective to proceed to human testing as rapidly as possible to test the hypothesis in the relevant species after proving basic, necessary pharmacological and toxicological features in preclinical models. To that end, different questions must be answered – safety, tolerability, dosing, effect size, biomarkers – not more basic experiments with disease models. It also means that additional attention and resources are needed to safely and ethically facilitate the human-tissue and human experience-based data that often fall under the rubric of "translational research" (doi: 10.1126/scitranslmed.aaa2049). The traditional world of "self-regulated" fundamental and applied research is dominated by academic research, while later testing of putative vetted therapeutic candidates is performed primarily by industry, with some academic contributions to industry-sponsored clinical trials or investigator-initiated trials. New advanced therapies (i.e., cell and gene therapies, tissue engineering, MedTech and their combinations) has closed the gap between the regulatory world and the academic world as many clinical developments in this field are driven by academic labs, with many new regulatory challenges. If this trend is to continue, then educating young researchers with a "mind-set" targeted towards the basics of translational medicine and important

regulatory realities will also be necessary (doi: 10.1126/scitranslmed.aaa0599). In addition, the bridge between translational project idea-generating fundamental research and the subsequent regulated clinical development can accelerate and de-risk the translation if high standards of quality are applied in this transition phase.

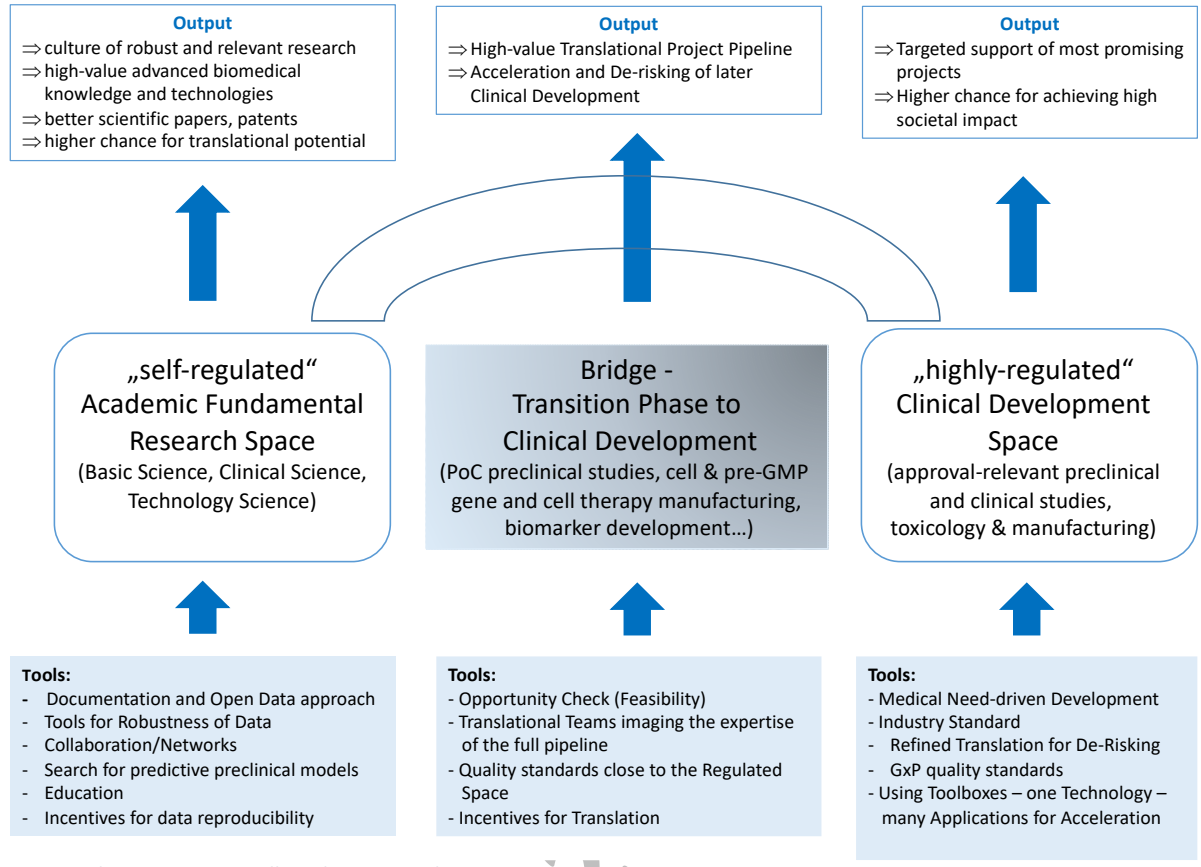
Reproducibility, veracity and validity of early mechanistic therapeutic data comprise the foundation upon which the entire edifice of clinical research is built. Today the success rate of translation is woefully low – even among drugs that enter phase I human trials, fewer than 10% are registered as new drugs (Deloitte Center for Health Solutions, Ten Years On: measuring the return from pharmaceutical innovation 2019, https://www2.deloitte.com/uk/en/pages/life-sciences-and-healthcare/articles/measuring-return-from-pharmaceutical-innovation.html). Remarkably, the majority of product candidates failed not in early phase I/IIa trials (because of safety issues) but in later stage phase IIb or phase III trials (because of efficacy issues). As preclinical models have recognized limitations as discussed above, we must learn as much as possible in early clinical trials using accompanying mechanistic studies (safety, PD/PK, mode-of-action, surrogate markers). Reliable results from those studies require highly validated biomarker validation and their use in tests – another important standardization task to improve quality that is frequently underestimated.(https://doi.org/10.1038/s41584-018-0005-9) Lessons learned from those studies allow iterative improvements of the product candidate or patient selection by a back-to-bench-forward-to-bed approach ("refined translation") - an important de-risking process.

Developing a generally accepted set of best practices for ensuring that future success is limited only by biological constraints, rather than by methodological issues, is an important advance benefitting the entire biomedical translational research effort to more effectively and reliably address compelling unmet patient needs.

Additional issues to consider/expand:

- ***Utility of models depends on the specific disease: e.g., MS vs Alzheimer's, how much is already known about the mechanism, and how the model accurately mimics select, or relevant mechanistic features known in humans etc***
    - **Access to "correct" samples types: validating**
- ***Translational failures provide a currently untapped source of evidence (excluding failure due to "bad science") that should be exploited for studying data reproducibility using past experiences.***
    - **"consistent" reliable data are expected, and data that do not fit expectations are either gradually discredited and faded out, or interpreted to the "disadvantage" of the hypothesis**
- ***Aesthetic article format consideration and to reduce the abstract nature of the discussion:***
    - **Produce Text Boxes with examples for certain concepts e.g. biological differences between animal models and humans, hallmarks of bias, statistical no-nos, external vs internal validity.**

182 **Figure proposal –** *Three phases towards reproducibility in biomedical translation*



183
184 *Figure credit Hans-Dieter Volk and Petra Reinke*