Preclinical Assessment for Translation to Humans (PATH): An Approach for Assessing Supporting Evidence for Early Phase Trials and Innovative Care

Jonathan Kimmelman*[1], Patrick Kane[1], Selin Bicer[1], Benjamin Gregory Carlisle[1]

* Corresponding author
ph: 514 398 3306
email: Jonathan.kimmelman@mcgill.ca

Dept. of Equity, Ethics and Policy, Rm 1155
School of Population and Global Health
McGill University
2001 McGill College
Montreal, QC H3A 1L7
Canada

Word Count: 3484 (main text); 381 (box); 197 (abstract)

eTOC blurb: The ethical and scientific basis of early phase trials rests on supporting evidence from biochemical, preclinical and clinical studies. By parsing this evidence into nine mechanistic steps, scientists can communicate the strength of support for a trial in a manner that is comprehensive, transparent, and accurate.

Summary

Early phase trials and innovative care draw support from basic science, preclinical studies, and clinical research. Such evidential diversity presents a challenge for traditional ways of synthesising evidence. In what follows, we review the limitations of existing approaches for communicating supporting evidence for early phase trials. We then offer a structured approach, PATH (Preclinical Assessment of Translation to Humans). PATH is grounded on the premise that the case for administering novel strategies to patients requires connecting the dots between nine mechanistic steps supporting a clinical claim. Using PATH entails first parsing supporting evidence, assessing the strength of evidence at each step, and then assessing the strength of a chain of evidence linking drug administration to clinical effect. While PATH requires further refinement, the approach reduces some of the opacity, arbitrariness, and bias in current ways of presenting and assessing scientific support for early phase trials and innovative care.

Introduction

Most new drugs, vaccines or devices are evaluated in large, randomized trials before uptake in clinical practice. To reach that point, new treatments must first be tested in early phase trials, including phase 1 trials (which evaluate safety and dosing) and phase 2 trials (which gather preliminary evidence of efficacy). Sometimes, innovative care is provided to patients absent any support from trials.

Both early phase trials and innovative care must meet some threshold for probability of success. Absent that, they are unlikely to redeem toxicities or use resources efficiently[1]. However, neither early phase trials nor innovative care can draw on the strong forms of evidence that are used to support other decisions in medicine. Indeed, their justification rests on forms of evidence like mechanism studies[2], preclinical studies, or case reports, each of which involve risk of bias and threats to reproducibility[3–7].

Various episodes suggest many early phase trials are launched without good reason to expect the drug will ultimately translate[8–16]. In one example, a healthy volunteer died and several others were injured after participating in a phase 1 trial that was later faulted for lacking preclinical support[9]. Other reasons to think some early phase trials rest on fragile supporting evidence include repeated observations that most preclinical studies fail replication, and that scientific justifications in protocols or publications of early phase trials often have gaps[17,18]. For innovative care, the abundance of clinics offering non-validated cell-based interventions for a variety of disorders[19] suggest that support for novel treatment protocols is often weak.

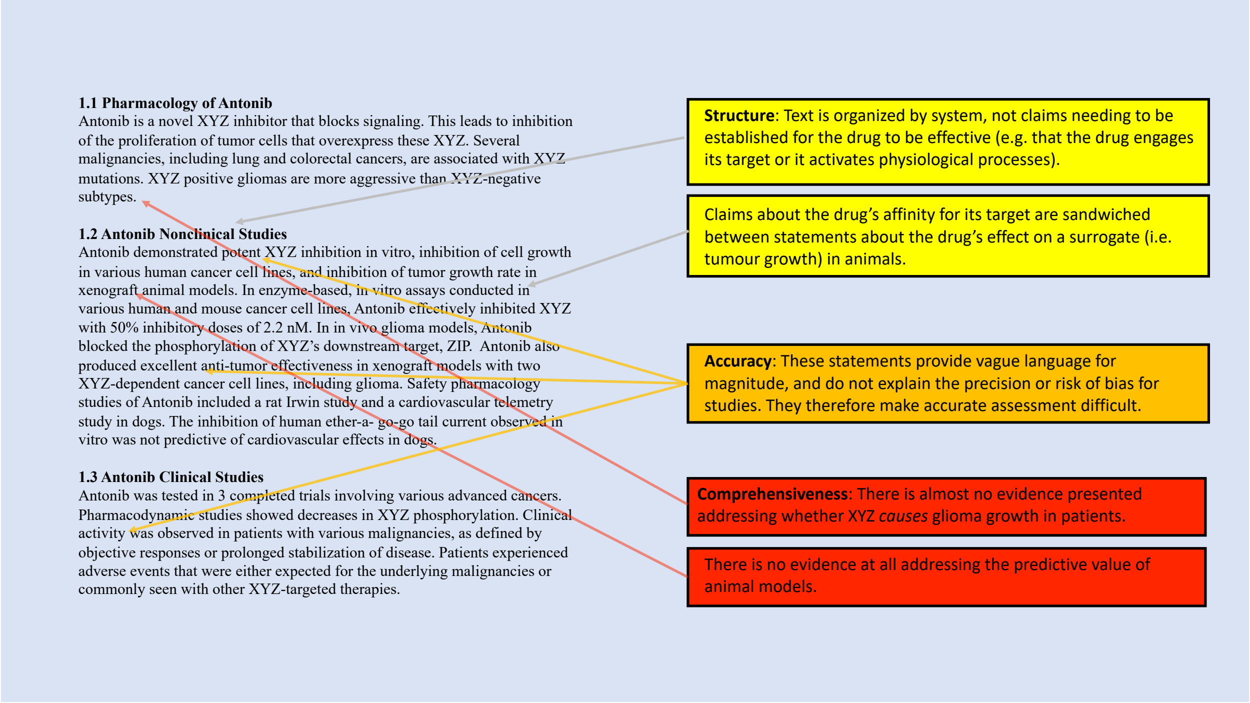
Despite the importance of a sound scientific rationale, formal approaches for building or assessing evidence supporting early phase trials or innovative care are lacking. This limits the ability of scientists, investigators, physicians and/or oversight committees to safeguard patients and maximize the impact of trying novel approaches for the first time in patients.

In what follows, we offer a structured approach, PATH (Preclinical Assessment for Translation to Humans), that scientists can use to communicate the scientific rationale for early phase trials or innovative care. Similarly, physicians or oversight bodies can use the approach to assess novel treatment strategies.

Current Standards for Presenting or Assessing Supporting Evidence

Any approach to justifying a trial or novel care approach should meet four criteria. First, the approach - like others for assessing strength of evidence[20] - should be structured. Structured approaches support efficient workflows and integrating different considerations into a decision. Second, the approach should be comprehensive in capturing relevant evidence. That way claims of clinical promise can draw on diverse methodologies and findings. Third, the approach should foster accurate judgments about the strength of evidence. Judgments about the risk, benefit and scientific value of a trial should scale with the number of supporting studies, their rigour and the extent to which they address key gaps. Last, the approach should encourage transparency. That way, scientists and assessors can explain the basis for their judgments, and, when needed, isolate reasons for disagreement.

The way trial protocols are currently written makes meeting these four criteria almost impossible. In **Figure 1**, we excerpt supporting evidence from a trial protocol (details are masked and text shortened for brevity). While protocols often contain narratives that make them easy to read, they often lack structures or information that support informed, critical and integrated appraisals.



1.1 Pharmacology of Antonib
Antonib is a novel XYZ inhibitor that blocks signaling. This leads to inhibition of the proliferation of tumor cells that overexpress these XYZ. Several malignancies, including lung and colorectal cancers, are associated with XYZ mutations. XYZ positive gliomas are more aggressive than XYZ-negative subtypes.

1.2 Antonib Nonclinical Studies
Antonib demonstrated potent XYZ inhibition in vitro, inhibition of cell growth in various human cancer cell lines, and inhibition of tumor growth rate in xenograft animal models. In enzyme-based, in vitro assays conducted in various human and mouse cancer cell lines, Antonib effectively inhibited XYZ with 50% inhibitory doses of 2.2 nM. In in vivo glioma models, Antonib blocked the phosphorylation of XYZ's downstream target, ZIP. Antonib also produced excellent anti-tumor effectiveness in xenograft models with two XYZ-dependent cancer cell lines, including glioma. Safety pharmacology studies of Antonib included a rat Irwin study and a cardiovascular telemetry study in dogs. The inhibition of human ether-a- go-go tail current observed in vitro was not predictive of cardiovascular effects in dogs.

1.3 Antonib Clinical Studies
Antonib was tested in 3 completed trials involving various advanced cancers. Pharmacodynamic studies showed decreases in XYZ phosphorylation. Clinical activity was observed in patients with various malignancies, as defined by objective responses or prolonged stabilization of disease. Patients experienced adverse events that were either expected for the underlying malignancies or commonly seen with other XYZ-targeted therapies.

Structure: Text is organized by system, not claims needing to be established for the drug to be effective (e.g. that the drug engages its target or it activates physiological processes).

Claims about the drug's affinity for its target are sandwiched between statements about the drug's effect on a surrogate (i.e. tumour growth) in animals.

Accuracy: These statements provide vague language for magnitude, and do not explain the precision or risk of bias for studies. They therefore make accurate assessment difficult.

Comprehensiveness: There is almost no evidence presented addressing whether XYZ *causes* glioma growth in patients.

There is no evidence at all addressing the predictive value of animal models.

Figure 1: A representative description of supporting evidence for an early phase trial. In the above example, XYZ is a hypothetical gene that is believed to play a key role in cell growth and migration; Hyperactivation triggers a network of interrelated signaling pathways promoting cancer growth. Antonib is a hypothetical drug designed to inhibit XYZ activation. The above text is broadly representative of the kind of narrative review of evidence used to support early phase trials. Deficiencies in terms of structure, comprehensiveness and accuracy are indicated.

Existing Approaches for Preclinical Evidence Synthesis

Before developing PATH, we conducted a scan of various methods that are used- or that could be used- for evaluating supporting evidence in early phase trials. Though a systematic review of such approaches would be valuable, we did not opt for this approach owing to the many philosophical judgments that would need to be made for designing literature searches, deciding which approaches would be applicable, and coding the content of such documents.

The approaches we identified fall short on one or more of the criteria above. Various groups and scientific societies[21–29] have offered approaches for assessing preclinical experiments (i.e. testing a drug's effect in live animals). A more formal and influential approach has adapted the GRADE framework (used to assess strength of evidence in clinical medicine[30]) for rating the certainty of evidence for preclinical studies[31,32]. While these approaches are valuable, they have philosophical and pragmatic shortcomings for the task at hand. For example, they assign a secondary role for using mechanistic evidence. One document states "the GRADE framework does not explicitly address mechanistic data, but they may be used to inform judgments about indirectness."[33] The approaches provide minimal to no guidance for assessing mechanistic evidence, when it is used, or for assessing the predictive value of models. And yet much of the evidence used to support early phase trials involves discrete claims about pathophysiology. Such approaches are therefore not strong on comprehensiveness and accuracy. GRADE and other approaches above encourage a careful, study by study assessment of strength of evidence. However, they do not disarticulate and reassemble various constituent claims in a trial protocol such that key evidence gaps can be spotted. Another limitation of approaches like GRADE is that they are mainly aimed at decision-makers evaluating strength of evidence. However, they provide little prospective guidance for researchers seeking to assemble evidence to support an early phase trial or innovative care protocol.

Regulators like the EMA[34,35] and FDA[36] offer guidance for preclinical evidence supporting early phase trials (the latter's are only for gene and cell therapy). Though fairly comprehensive, these guidelines do not offer much structure. The EMA guidance does not address reproducibility threats, for example.

Emmerich et al offers a "critical path" approach for target assessment[37]. Some pharmacologists have also offered similar critical path-like approaches for selecting starting doses for early phase trials[38]. These promote accuracy by pointing to validity threats and by decomposing information needed to support trial initiation, but they generally focus on only one aspect of trial decision-making (i.e. target assessment).

Preclinical Assessment of Translation to Humans (PATH)

PATH builds on recent work in philosophy of evidence-based medicine that has sought to restore a role for mechanism in evaluating the strength of evidence[39,40]. The main point of this literature is that traditional approaches to evidence-based medicine tend to deprioritize mechanistic evidence. However, knowledge of mechanism plays a key role in formulating judgments about the strength of a scientific claim. For example, parachutes have never been tested in randomized trials as a prophylactic against gravitational challenge[41]. Yet we have strong grounds, rooted in mechanism, for confidence in their efficacy. In many realms of medicine, most of the evidence for asserting a claim of efficacy or effectiveness rests primarily on mechanistic evidence. This is certainly the case in early phase research. Rather than understanding mechanism as “evidence of last resort,” newer approaches to evidence-based medicine regard mechanism as complementary to evidence from well-designed trials.

To develop PATH, we started by parsing the clinical effects of a drug into four steps that are common targets for separate investigations. The aim of experiments in model systems is to mirror these four steps in systems that have some predictive value for the clinical scenario. After developing the basic PATH approach, we sought review from a diverse panel of experts in evidence and early-stage translational research (see **Appendix**). We modified PATH based on these comments.

The PATH approach offers a pragmatic device for helping experts organize, integrate, and interpret various evidentiary claims. It is not intended as a complete description of what is going on in experimental and target systems. PATH is founded on the premise that the case for pursuing early phase trials or innovative care requires connecting the dots from evidence of mechanism to clinical outcomes. Throughout this essay, we will illustrate our approach using a hypothetical trial of a novel XYZ inhibitor, “Antonib” that is being tested against glioma.

The main task for assessing a proposal is to determine the level of confidence that a drug, when given to a target population, will produce a desired response (hereafter we call this the “target scenario”). In our example, the target scenario is that Antonib given to glioma patients will improve survival. The higher the confidence in efficacy in a target scenario, the more solid the ethical and scientific rationale for the endeavor[42,43].

Though the evidence used to support initial attempts at a novel therapy assumes a dizzying variety, supporting evidence can be parsed into nine mechanistic steps that, when pieced together, build a chain of evidence supporting the drug’s clinical promise.

In **Figure 2**, we offer a PATH diagram that captures these nine steps. The left half of the diagram (“direct steps”) reflects mechanistic processes that need to occur for the drug to be effective in target patients. The right half (“model steps”) concerns parallel mechanistic processes in model systems (e.g. cell culture, animal experiments, or even trials of the same drug in related diseases). Horizontal arrows (“translational steps”) connect findings in model systems to target scenarios. Almost every piece of supporting evidence described in a typical protocol can be assigned to at least one of the steps in a PATH diagram.

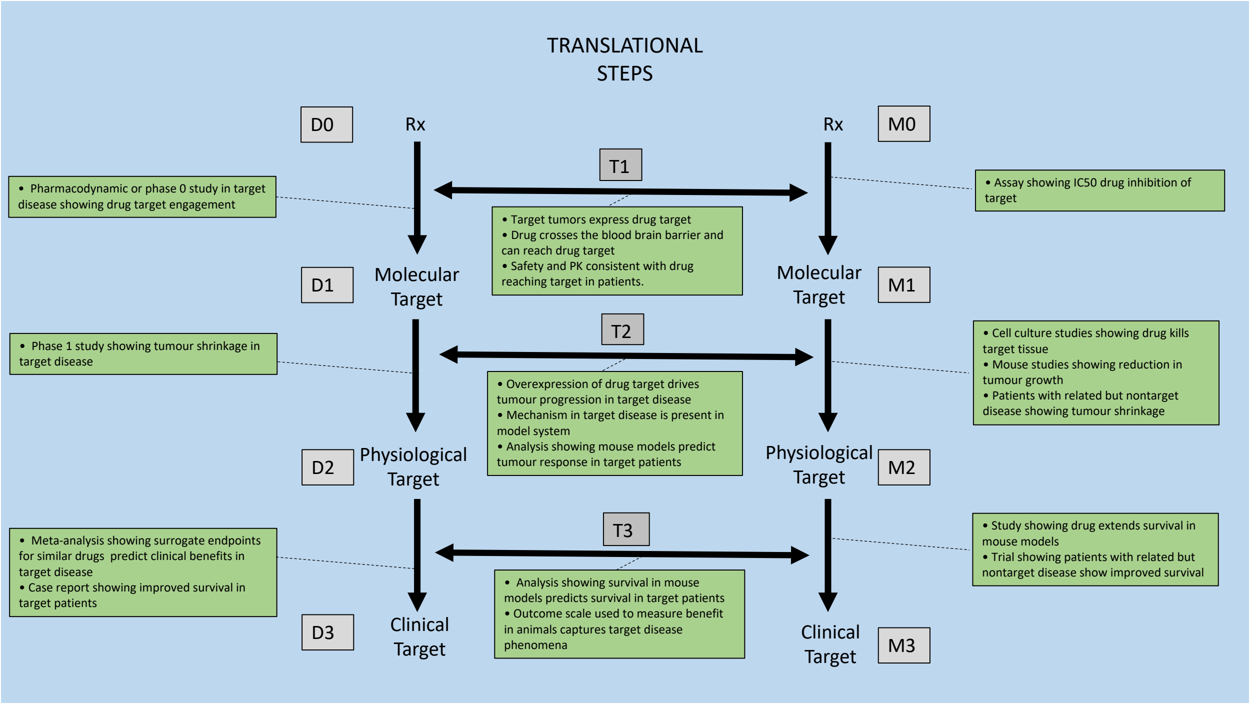

TRANSLATIONAL STEPS
D0
Rx
Rx
M0
T1
• Pharmacodynamic or phase 0 study in target disease showing drug target engagement
• Target tumors express drug target
• Drug crosses the blood brain barrier and can reach drug target
• Safety and PK consistent with drug reaching target in patients.
• Assay showing IC50 drug inhibition of target
D1
Molecular Target
Molecular Target
M1
T2
• Phase 1 study showing tumour shrinkage in target disease
• Overexpression of drug target drives tumour progression in target disease
• Mechanism in target disease is present in model system
• Analysis showing mouse models predict tumour response in target patients
• Cell culture studies showing drug kills target tissue
• Mouse studies showing reduction in tumour growth
• Patients with related but nontarget disease showing tumour shrinkage
D2
Physiological Target
Physiological Target
M2
T3
• Meta-analysis showing surrogate endpoints for similar drugs predict clinical benefits in target disease
• Case report showing improved survival in target patients
• Analysis showing survival in mouse models predicts survival in target patients
• Outcome scale used to measure benefit in animals captures target disease phenomena
• Study showing drug extends survival in mouse models
• Trial showing patients with related but nontarget disease show improved survival
D3
Clinical Target
Clinical Target
M3

Figure 2: PATH diagram. In the PATH approach, the goal is to trace a chain of evidence from administration of the drug in a target scenario (D0) through to a desired clinical effect (D3). For early phase trials and innovative care, other early phase trials or clinical experience may supply some direct (D) evidence of drug efficacy. However, much of the evidence creating a chain will derive from studies in model systems (M) coupled to evidence of translational (T) relevance. Text boxes provide examples of evidence that might support different steps.

To assert potential efficacy in the target scenario, evidence within protocols must suggest at least one path through mechanistic steps, from administering the drug to achieving efficacy in target patients. At the point of early phase trials, direct steps on the left half of the PATH diagram can be only weakly substantiated using evidence from case reports or small trials. Sponsors must supplement this evidence using studies involving models and evidence of the relevance of those models.

Direct and Model Steps

Direct evidence for a drug's promise typically comes from case reports or phase 0 or 1 studies. Model evidence derives from biochemical studies, in vitro experiments, preclinical studies and clinical trials.

Though there are an infinite number of mechanistic steps between treatment and effect in target patients, the chain of events from drug administration to clinical effect can be described as being mediated through four major steps: administering a treatment (D0 or M0), engaging a drug target (D1 or M1), altering a pathophysiological process (D2 or M2), and producing a clinical response (D3 or M3). This parsing of mechanistic steps reflects that drug developers generally build their efforts around a drug's target, and use mediating physiological processes (e.g. for cancer, tumour shrinkage) as surrogates for clinical benefits like survival.

In our example, researchers might establish that Antonib inhibits XYZ *in vitro* (M0 to M1). They also might establish that XYZ inhibition leads to tumour growth inhibition in mice (M1 to M2), and that tumour growth inhibition leads to greater survival in those mice (M2 to M3).

Translational Steps

Translational evidence is needed to connect model systems to target scenarios. This entails evidence that mechanisms driving the target clinical disorder are recapitulated in model systems.

Evidence supporting translational steps often take three obvious forms. The first is evidence that model systems recapitulate mechanistic processes driving target disease[44]. Our Antonib protocol might provide evidence that XYZ activates a signal transduction cascade that leads to progression of glioma, and that this complete set of processes is present in the model.

Second, translational steps can be supported by providing evidence that animal models, intervention doses or outcomes used in model systems reflect the target scenario ("construct validity"[45]). Several types of studies might be invoked for this. Toxicology, safety, and pharmacokinetic studies, for example, are essential to establishing the construct validity of interventions in preclinical studies, since efficacy in models is unlikely to carry over to humans if drugs are used at intolerable doses.

Third, translational steps can be supported by explaining that effects of the drug have been observed in several different model systems. Such statements provide evidence that cause and effect relationships in model systems are robust against changes in context ("external validity"[46]). Other more subtle forms of translation evidence are described in Box 1.

Applying PATH

Generating the case for attempting a novel strategy (or the assessment thereof) occurs in three stages. First, supporting evidence is identified and assigned to steps in the PATH diagram. Second, the level of confidence for each step is assessed. Third, a cumulative judgment is rendered about the prospects that the intervention will show efficacy in the target scenario. In what follows, we describe how a sponsor might build the argument to initiate an early phase trial. A similar approach, however, might be used by a reviewer to retrospectively adjudicate that argument (in the **Appendix**, we offer an example of how PATH might be used by a reviewer to analyze evidence presented in a phase 2 trial protocol we accessed on ClinicalTrials.gov).

Stage 1: Assigning Evidence to Steps

Those making the case for an early phase trial or innovative care should begin by using the PATH diagram to systematically search for evidence supporting relevant steps. Assigning direct or model evidence to steps in the PATH diagram is generally straightforward. Any study testing the effects of a treatment in the target population, like a prior early phase trial, contributes evidence to direct steps. Studies testing a treatment in a different but related disease, or in animal models, supply evidence for model steps. In some cases, a piece of evidence might skip steps: an experiment might show a drug leads to tumour inhibition in mice without, in the same experiment, showing evidence of target engagement (M0 to M2).

Evidence can be assigned to translational steps if it addresses the relationship between models and target scenarios at relevant levels in the PATH diagram. Various types of evidence that might be used to support translational steps are offered in **Box 1**.

Where possible, citations should be provided for each piece of evidence so that mechanistic claims can be investigated further.

Box 1: Types of Evidence for Translational Steps

Evidence substantiating a model system's predictive value for a target scenario often takes the following five forms.

The first is evidence supporting target disease pathophysiology (and along with it, reasons to believe such pathophysiology is recapitulated in model systems). This can establish that target inhibition within *in vitro* studies, or disruption of a pathophysiological process in an in vivo system will translate to a disease response in the target scenario. Examples of evidence taking this form would be a body of basic science research showing the role of some molecular process in driving human disease (e.g. papers showing XYZ overexpression drives glioma).

The second is an explanation of relationships between various features of a model systems and those for the target scenario ("construct validity)[47]. For example, sponsors might address a) interventions used in model systems (e.g. are doses used in models representative of those tolerated in patients?); b) populations (e.g. do animal models recapitulate essential aspects of human pathophysiology?); and c) outcomes (e.g. do outcome measures, like performance on a rotarod test, provide a read-out of a human disease phenomena like Parkinsonism?).

A third line is replication of effects in different model systems ("external validity")[46]. For example, researchers might assert that their drug shows large effects in three different model systems. This suggests a robust causal process that has greater prospects of withstanding translation to a target scenario.

Fourth is evidence suggesting the absence of "interfering effects" in the target scenario. Mechanisms that operate in both model and target scenarios can be attenuated in the latter if there is some other mechanism or context in the target system that counteracts an intervention's effects. This might happen because a drug activates some other process that buffers the drug's effect on a clinical outcome (e.g. development of resistance, or activation of metabolic processes that limits exposure to a drug). Or it might occur because a drug has intolerable side effects that lead to nonadherence, thus interfering with the translation of preclinical findings. Studies showing the absence of such interference can reinforce the specificity of translational claims.

The fifth evidence form is a systematic review of a model's predictive value[47]. Using meta-analysis, how well did preclinical studies using a particular system predict drug efficacy in patients?

Stage 2: Assessing the Strength of Evidence at Each Step

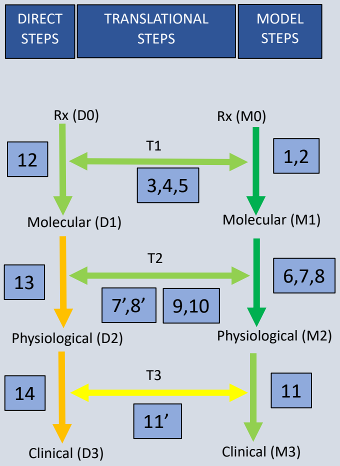
The next task is to assess the level of confidence for each step. With direct or model steps, this is facilitated by considering three factors, where possible. The first is magnitude. The second is the precision of the effect estimate. Information about precision might take various forms, including inferential statistics, confidence intervals around effects, or even simple statements about sample size). The third factor is risk of bias. The techniques for protecting internal validity in model experiments will be similar to those used for trials[40], including use of randomization, prospective registration, and blinded outcome assessment[45] (see elsewhere[48,49] preclinical risk of bias assessment tools). For all other forms of evidence, measures for reducing risk of bias will depend on the experimental techniques.

Assessing the level of confidence for translational steps is more complicated. However, it is often possible to assess this evidence in terms of the magnitude of relationships connecting models to target scenarios, as well as precision and risk of biases[50]. For example, sponsors frequently assert that certain cancers are driven by an oncogene based on the level of the oncogene's expression in resected tumours. However, such evidence is susceptible to confound because it is derived from observational studies. Confidence would be greater if such claims derived from experimental manipulations of resected tissues. The diagnostic concepts of specificity and sensitivity can also be useful for assessing models. Further information on assessing the level of confidence for translational steps is offered in **Table 1**.

To synthesize judgments, PATH diagrams can be colour coded at each step[40], using green to designate high confidence steps and red to indicate an absence of supporting evidence. This can facilitate the next stage for PATH.

Stage 3: Combining Judgments About Strength of Evidence

At this point, evidence supporting a novel therapeutic strategy can be represented with an annotated PATH diagram (see **Figure 3**). This information can then be used to assess support for the overall proposal.



TARGET MODULATION IN MODELS (M0-M1)

1. Antonib IC50 inhibition of XYZ is 2.2 nM.
2. In randomized xenograft glioma models (see #6 below), Antonib led to 157% lowered expression of a ZIP, a downstream target of XYZ.

TRANSLATABILITY (T1)

3. Antonib crossed the blood brain barrier at a rate that was 20% more efficient than morphine.
4. 70% of tumours resected from glioma patients show of XYZ overexpression in blinded pathology studies.
5. Doses needed to achieve target inhibition in models, when scaled to patients, did not cause any SAEs in studies involving 20 patients.

PHYSIOLOGICAL EFFECT IN MODELS (M1-M2)

6. Antonib given to 8 of 9 different types of glioma cell lines showed growth inhibition of 60% or more.
7. Randomized studies of Antonib in glioma-bearing orthotopic mice significantly shrank by 50% compared with temazolomide. Results were reproduced three times.
8. Antonib given to patients with lung or colorectal cancers resulted in ≥ 40% objective response (tumour assessments were blinded).

TRANSLATABILITY (T2)

7'. Glioma-bearing orthotopic mice are sensitive, but not specific models for human glioma. Doses in this study were similar to those planned for this trial

8'. Lung and colorectal cancers are driven by XYZ over-activation. Doses used in the above (and present) trial were tolerable (5% of patients experienced serious adverse events)

9. Glioma progression is driven by XYZ hyperactivation, as indicated by several independent studies showing knock-down of XYZ in glioma cells inhibits growth by 70% ($p < .05$).

10. Robustness of Antonib anti-tumour activity is suggested by achievement of significantly reduced tumour growth (>50% shrinkage) in randomized xenograft studies of three different cancers types.

CLINICAL EFFECT IN MODELS (M2-M3)

11. Antonib significantly doubled survival in randomized, blinded studies of glioma xenograft mice.

TRANSLATABILITY (T3)

11'. Doses of Antonib in #10 matched those planned for this trial. The model is very sensitive but not very specific for predicting human response.

TARGET MODULATION IN THE TARGET SCENARIO (D0-D1)

12. Antonib significantly doubled XYZ phosphorylation in gliomas resected from patients in a phase 0 trial.

PHYSIOLOGICAL EFFECT IN THE TARGET SCENARIO (D1-D2)

13. Antonib produced a partial response in one of two glioma patients in a phase 1 trial involving mixed cancers.

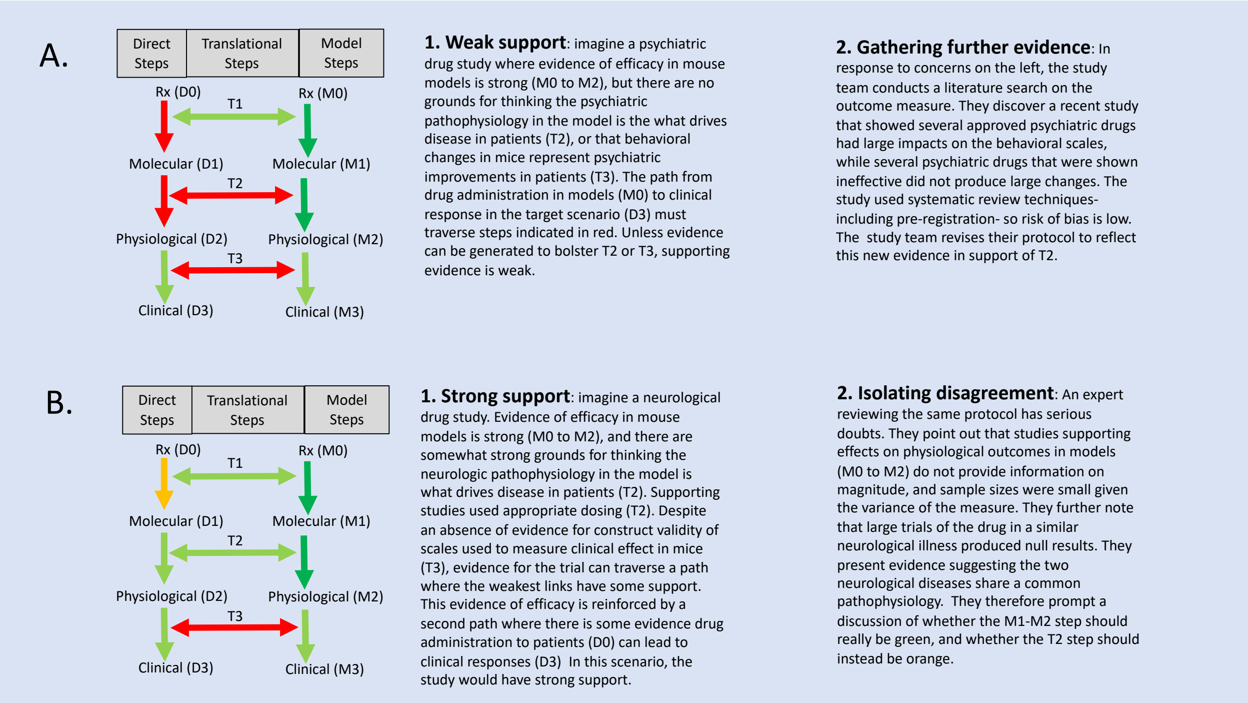
CLINICAL EFFECT IN THE TARGET SCENARIO (D2-D3)

14. A meta-analyses of glioma trials showed tumour response was moderately but significantly correlated with increased overall survival (R^2=0.4).

Figure 3: A PATH description of supporting evidence. The contents of Box 1 are displayed with an annotated PATH diagram (with supplementation of missing information). Colour coding represents the assessor's judgment about strength of evidence (red = weak support; yellow = equivocal support; solid green = strong support). Numbers in PATH diagram correspond to evidence presented in juxtaposed text; superscripted primes indicate construct validity information provided in studies assigned to model evidence.

What is key is that there is at least one chain of evidence linking administration of the drug (at M0 or D0) to clinical effects in the target scenario (D3). Not all steps need to be populated for evidence to be strong. Also, there may be cases where a chain of evidence skips over a direct or model step (for example, where there are no measures of physiological outcomes). That does not necessarily undermine the overall confidence in the strategy, provided there is evidence supporting translation from models to target scenarios at the clinical level.

Assertions of a drug's value in the target scenario will often be limited by the PATH steps that have the weakest support for a chain of evidence (see **Figure 4**). Scientists might use additional studies to fortify a weak link in a chain of evidence. Alternatively, they might seek evidence that supplements a weak chain by building another chain in the PATH diagram.



1. Weak support: imagine a psychiatric drug study where evidence of efficacy in mouse models is strong (M0 to M2), but there are no grounds for thinking the psychiatric pathophysiology in the model is the what drives disease in patients (T2), or that behavioral changes in mice represent psychiatric improvements in patients (T3). The path from drug administration in models (M0) to clinical response in the target scenario (D3) must traverse steps indicated in red. Unless evidence can be generated to bolster T2 or T3, supporting evidence is weak.

2. Gathering further evidence: In response to concerns on the left, the study team conducts a literature search on the outcome measure. They discover a recent study that showed several approved psychiatric drugs had large impacts on the behavioral scales, while several psychiatric drugs that were shown ineffective did not produce large changes. The study used systematic review techniques- including pre-registration- so risk of bias is low. The study team revises their protocol to reflect this new evidence in support of T2.

1. Strong support: imagine a neurological drug study. Evidence of efficacy in mouse models is strong (M0 to M2), and there are somewhat strong grounds for thinking the neurologic pathophysiology in the model is what drives disease in patients (T2). Supporting studies used appropriate dosing (T2). Despite an absence of evidence for construct validity of scales used to measure clinical effect in mice (T3), evidence for the trial can traverse a path where the weakest links have some support. This evidence of efficacy is reinforced by a second path where there is some evidence drug administration to patients (D0) can lead to clinical responses (D3) In this scenario, the study would have strong support.

2. Isolating disagreement: An expert reviewing the same protocol has serious doubts. They point out that studies supporting effects on physiological outcomes in models (M0 to M2) do not provide information on magnitude, and sample sizes were small given the variance of the measure. They further note that large trials of the drug in a similar neurological illness produced null results. They present evidence suggesting the two neurological diseases share a common pathophysiology. They therefore prompt a discussion of whether the M1-M2 step should really be green, and whether the T2 step should instead be orange.

Figure 4. Completing a mechanistic path to clinical effect. Two scenarios illustrate how PATH diagrams can be used to map mechanistic steps towards clinical effects. The scenarios also illustrate how PATH can be used to improve the preparation and/or discussion of supporting evidence. In scenario A, researchers prepare a PATH diagram, and find it wanting on a crucial step. This provides an occasion to gather more evidence to strengthen support for the trial. In scenario B, a research team presents a PATH diagram that seems to provide strong support. Note that there are two paths from drug administration to clinical effect involving at least moderately strong evidence at each step. The first is from M0 to M1, with a T1 step leading to achievement of D1. The second is from M0 to M1 to M2, with a T2 step leading to achievement of D2. The fact that there are two moderately strong paths bolsters the case for this drug. However, a reviewer has raised concerns about the precision, bias and inconsistent findings from model studies. These concerns now provide a basis for focusing discussions on the strength of evidence for the intervention.

Discussion

PATH offers a general approach for presenting evidence that a new treatment will deliver on its promise (alternatively, it can be used in review to assess the strength of a trial proposal). By parsing evidence into nine mechanistic steps and showing how they articulate, it offers structure. By soliciting evidence for each step, including translational claims, it encourages comprehensiveness. PATH fosters accuracy by encouraging consideration of magnitude, precision, and risk of bias. It also fosters accuracy by exploiting the value of decomposing complex questions (i.e. whether a drug will eventually prove effective) into smaller sub-questions (i.e. whether a drug engages its target) and reassembling them. Decomposition has been shown to foster more accurate judgments where uncertainty is high[51]. Finally, PATH promotes transparency: when assessors disagree, annotated PATH diagrams can help scientists isolate sources of disagreement, thus centering discussions and further research.

The PATH approach does, however, have limitations that necessitate further work and/or extension. Here, we lay out what we regard as the six most important ones, and possible avenues for addressing each.

First, some might regard PATH as impractical. Clinical investigators may chafe at presenting supporting evidence in a new format, and IRBs might find PATH overly exacting. These downsides must be considered against the value of encouraging more rigorous and transparent processes. These concerns might be addressed in future work by creating protocol templates for investigators and using software for creating PATH diagrams.

A second limitation is that presenting evidence for translational edges is difficult. Others offering approaches for assessing preclinical evidence have similarly struggled with this challenge (see, for example, Hooijmans et al[31]). Further work will be needed to develop a structure for presenting translational evidence.

Third, our approach does not provide a cookbook for constructing or evaluating supporting evidence. As noted, mechanistic and translational evidence draw on a variety of methodologies, each presenting different internal validity threats or notions of what constitutes a large effect size. However, our goal is not to erase the need for expertise, but rather to provide a scaffolding for its effective application.

Fourth, PATH should be understood as a way of mapping evidence. In doing so, PATH diagrams simplify or even distort complex causal processes to facilitate their critical engagement- much as a city metro map simplifies spatial relationships in a city to facilitate navigation. Each step in a PATH diagram could be exploded in greater detail or customized to specific contexts. For a vector-based gene therapy to work, the vector must transfect the appropriate cells and cause stable transgene expression before modifying a disease course. PATH diagrams for gene therapy trials might benefit by the addition of levels for transfection and gene expression. Also, PATH necessitates simplifying some types of evidence. A study showing that a drug causes disease response in animal models (M0-M2) would, for the sake of simplicity, be represented in the M1-M2 step of a PATH diagram.

Fifth, the assertion that PATH will improve transparency, accuracy, or evidence comprehensiveness is an empirical claim that is, perhaps ironically, grounded solely on mechanistic evidence. To achieve these ends, workflows for using PATH will need to be developed. These workflows will need to be tested in randomized trials against standard narrative evidence presentations to determine whether they in fact support more judgments that are more transparent, accurate and based on a comprehensive gathering of evidence.

Last, PATH will benefit from further expert input, refinement, and interpretation. This might be achieved by establishing working groups akin to those used to develop and refine GRADE. The present manuscript should be seen as an initial step in that process.

Conclusion

The justification for administering novel and unproven interventions- whether in research or care- rests in part on supporting evidence. Early phase trials and innovative care vary in the strength of evidence supporting them. Because supporting evidence takes many different forms, common techniques for synthesizing evidence and assessing its strength, like meta-analysis, have limited value. This greatly complicates the task of communicating and or assessing the promise of such endeavours.

Despite this variety of evidence, supporting evidence has a common basic structure. PATH aims at surfacing this structure to help scientists and others arrive at transparent judgments that reflect comprehensive use of evidence and accurate levels of confidence about achieving clinical goals. Absent a PATH-like approach, unstructured processes that prevail today are susceptible to arbitrary, opaque, and biased decisions.

END

Table 1: Assessing the Strength of Evidence for Translational Steps

Forms of Evidence	Examples of Statements addressing Translational Steps	Magnitude	Precision	Bias
Description of disease mechanism	Overactive XYZ expression suppresses apoptosis, thus promoting survival and growth of tumour tissue.	Extent to which various experiments point to the presence and influence of this mechanism as a driver of disease.	Number of conditions or settings where this purported mechanism has been observed to strongly drive outcomes	Are there reasons to think evidence supporting a mechanism is biased (e.g. negative studies not reported, or mechanistic evidence based on correlative evidence, not causal evidence?)
Description of Model Systems (Construct validity)	We delivered 0.10 mg Antonib to immunocompetent, genetically engineered mouse models of breast cancer.	NA (already assessed at vertical steps)	NA (already assessed at vertical steps)	Are there any ways the model experiments might over-predict benefit in a target (e.g. higher than tolerable doses used in models? animal model prone to exaggerated responses?)
Replication in Different Models (External validity)	Experiments in mice and pigs showed reduction in accumulation of plaque	Extent to which replication studies showed similarly meaningful effect sizes	Number of different replication studies	Is it possible some replication studies in models were negative, but not reported?
Absence of interfering effects	When given to patients, drug X does not trigger inflammatory processes that limit the ability of lymphocytes to tissues.	Extent to which experiments rules out the presence of an interferng mechanism	Number of observations where interfering mechanism not observed	Are there ways experiments might under-detect interfering mechanisms? Have researchers failed to address interfering mechanisms?
Evidence of Predictivity	Meta-analysis showed a strong linear relationship between effects of interactions in models and humans	Slope of correlation between effect sizes in models and target systems	Level of variance for strength of correlation	Did the meta-analysis use proper registration and inclusion criteria? Were risk of bias analyses performed?

Acknowledgments: This work was funded by NIH and CIHR. We particularly thank Jon Williamson, Veli-Pekka Parkkinnen and Scott Plotkin for critical feedback on earlier versions of PATH. We also thank referees who provided feedback on an earlier draft (see appendix), as well as those reviewing the submitted manuscript.

Declaration of Interests: JK serves on a data monitoring committee of NIAID, which reviews early phase trials; honoraria for this sum to less than $1000 USD/year. The authors declare no competing interests.

REFERENCES

1. Hay, M., Thomas, D.W., Craighead, J.L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. Nature Biotechnology *32*, 40–51. 10.1038/nbt.2786.

2. Vere, J., and Gibson, B. (2019). Evidence-based medicine as science. J Eval Clin Pract *25*, 997–1002. 10.1111/jep.13090.

3. Errington, T.M., Iorns, E., Gunn, W., Tan, F.E., Lomax, J., and Nosek, B.A. (2014). An open investigation of the reproducibility of cancer biology research. eLife *3*, e04333. 10.7554/eLife.04333.

4. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News *533*, 452. 10.1038/533452a.

5. Jones, S.P., Tang, X.-L., Guo, Y., Steenbergen, C., Lefer, D.J., Kukreja, R.C., Kong, M., Li, Q., Bhushan, S., Zhu, X., et al. (2015). The NHLBI-Sponsored Consortium for preclinicAl assESsment of cARdioprotective Therapies (CAESAR). Circulation Research *116*, 572–586. 10.1161/CIRCRESAHA.116.305462.

6. Pusztai, L., Hatzis, C., and Andre, F. (2013). Reproducibility of research and preclinical validation: problems and solutions. Nat Rev Clin Oncol *10*, 720–724. 10.1038/nrclinonc.2013.171.

7. Steward, O., Popovich, P.G., Dietrich, W.D., and Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. Experimental Neurology *233*, 597–605. 10.1016/j.expneurol.2011.06.017.

8. Temple, S., and Studer, L. (2017). Lessons Learned from Pioneering Neural Stem Cell Studies. Stem Cell Reports *8*, 191–193. 10.1016/j.stemcr.2017.01.024.

9. Report by the Temporary Specialist Scientific Committee (TSSC), "FAAH (Fatty Acid Amide Hydrolase)", on the Causes of the Accident during a Phase 1 Clinical Trial (2016).

10. Kolata, G. (2018). Harvard Calls for Retraction of Dozens of Studies by Noted Cardiac Researcher. The New York Times.

11. Smith, M.A. (2014). Lessons learned from adult clinical experience to inform evaluations of VEGF pathway inhibitors in children with cancer. Pediatric Blood & Cancer *61*, 1497–1505. 10.1002/pbc.25036.

12. Kordower, J.H. (2016). AAV2-Neurturin for Parkinson's Disease: What Lessons Have We Learned? In Gene Therapy for Neurological Disorders: Methods and Protocols Methods in Molecular Biology., F. P. Manfredsson, ed. (Springer), pp. 485–490. 10.1007/978-1-4939-3271-9_32.

13. Ainsworth, S., Menzies, S.K., Casewell, N.R., and Harrison, R.A. (2020). An analysis of preclinical efficacy testing of antivenoms for sub-Saharan Africa: Inadequate independent scrutiny and poor-quality reporting are barriers to improving snakebite treatment and management. PLoS Negl Trop Dis *14*, e0008579. 10.1371/journal.pntd.0008579.

14. Ginsberg, M.D. (2009). Current Status of Neuroprotection for Cerebral Ischemia Synoptic Overview. Stroke *40*, S111–S114. 10.1161/STROKEAHA.108.528877.

15. Cohen, D. (2018). Oxford TB vaccine study calls into question selective use of animal data. BMJ *360*, j5845. 10.1136/bmj.j5845.

16. Scott, S., Kranz, J.E., Cole, J., Lincecum, J.M., Thompson, K., Kelly, N., Bostrom, A., Theodoss, J., Al-Nakhala, B.M., Vieira, F.G., et al. (2008). Design, power, and interpretation of studies in the standard murine model of ALS. Amyotrophic Lateral Sclerosis *9*, 4–15. 10.1080/17482960701856300.

17. Wieschowski, S., Chin, W.W.L., Federico, C., Sievers, S., Kimmelman, J., and Strech, D. (2018). Preclinical efficacy studies in investigator brochures: Do they enable risk–benefit assessment? PLOS Biology *16*, e2004879. 10.1371/journal.pbio.2004879.

18. Pratte, M., Ganeshamoorthy, S., Carlisle, B., and Kimmelman, J. (2019). How well are Phase 2 cancer trial publications supported by preclinical efficacy evidence? International Journal of Cancer *145*, 3370–3375. 10.1002/ijc.32405.

19. Turner, L., and Knoepfler, P. (2016). Selling Stem Cells in the USA: Assessing the Direct-to-Consumer Industry. Cell Stem Cell *19*, 154–157. 10.1016/j.stem.2016.06.007.

20. Atkins, D., Best, D., Briss, P.A., Eccles, M., Falck-Ytter, Y., Flottorp, S., Guyatt, G.H., Harbour, R.T., Haugh, M.C., Henry, D., et al. (2004). Grading quality of evidence and strength of recommendations. BMJ *328*, 1490. 10.1136/bmj.328.7454.1490.

21. Vassal, G., Houghton, P.J., Pfister, S.M., Smith, M.A., Caron, H.N., Li, X.-N., Shields, D.J., Witt, O., Molenaar, J.J., Colombetti, S., et al. (2021). International Consensus on Minimum Preclinical Testing Requirements for the Development of Innovative Therapies For Children and Adolescents with Cancer. Mol Cancer Ther *20*, 1462–1468. 10.1158/1535-7163.MCT-20-0394.

22. Fisher, M., Feuerstein, G., Howells, D.W., Hurn, P.D., Kent, T.A., Savitz, S.I., and Lo, E.H. (2009). Update of the Stroke Therapy Academic Industry Roundtable Preclinical Recommendations. Stroke *40*, 2244–2250. 10.1161/STROKEAHA.108.541128.

23. Chamuleau, S.A.J., van der Naald, M., Climent, A.M., Kraaijeveld, A.O., Wever, K.E., Duncker, D.J., Fernández-Avilés, F., and Bolli, R. (2018). Translational Research in Cardiovascular Repair. Circulation Research *122*, 310–318. 10.1161/CIRCRESAHA.117.311565.

24. Kimmelman, J., and Henderson, V. (2015). Assessing risk/benefit for trials using preclinical evidence: a proposal. Journal of Medical Ethics *42*, 50–53. 10.1136/medethics-2015-102882.

25. Gurusamy, K.S., Moher, D., Loizidou, M., Ahmed, I., Avey, M.T., Barron, C.C., Davidson, B., Dwek, M., Gluud, C., Jell, G., et al. (2021). Clinical relevance assessment of animal preclinical research (RAA) tool: development and explanation. PeerJ *9*, e10673. 10.7717/peerj.10673.

26. ALSUntangled Group (2015). ALSUntangled: introducing The Table of Evidence. Amyotroph Lateral Scler Frontotemporal Degener *16*, 142–145. 10.3109/21678421.2014.987476.

27. Grigorian-Shamagian, L., Sanz-Ruiz, R., Climent, A., Badimon, L., Barile, L., Bolli, R., Chamuleau, S., Grobbee, D.E., Janssens, S., Kastrup, J., et al. (2021). Insights into therapeutic products, preclinical research models, and clinical trials in cardiac regenerative and reparative medicine: where are we now and the way ahead. Current opinion paper of the ESC Working Group on Cardiovascular Regenerative and Reparative Medicine. Cardiovascular Research *117*, 1428–1433. 10.1093/cvr/cvaa337.

28. Dib, N., Menasche, P., Bartunek, J.J., Zeiher, A.M., Terzic, A., Chronos, N.A., Henry, T.D., Peters, N.S., Fernández-Avilés, F., Yacoub, M., et al. (2010). Recommendations for successful training on methods of delivery of biologics for cardiac regeneration: a report of the International Society for Cardiovascular Translational Research. JACC Cardiovasc Interv *3*, 265–275. 10.1016/j.jcin.2009.12.013.

29. Charis Wong, Jenna M. Gregory, Jing Liao, Kieren Egan, Hanna M. Vesterinen, Aimal Ahmad Khan, Maarij Anwar, Caitlin Beagan, Fraser Brown, John Cafferkey, et al. (2022). A Systematic Approach to Identify Neuroprotective Interventions for Motor Neuron Disease. medRxiv, 2022.04.13.22273823. 10.1101/2022.04.13.22273823.

30. Guyatt, G.H., Oxman, A.D., Vist, G.E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H.J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ *336*, 924–926. 10.1136/bmj.39489.470347.AD.

31. Hooijmans, C.R., Vries, R.B.M. de, Ritskes-Hoitinga, M., Rovers, M.M., Leeflang, M.M., IntHout, J., Wever, K.E., Hooft, L., Beer, H. de, Kuijpers, T., et al. (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLOS ONE *13*, e0187271. 10.1371/journal.pone.0187271.

32. Morgan, R.L., Thayer, K.A., Santesso, N., Holloway, A.C., Blain, R., Eftim, S.E., Goldstone, A.E., Ross, P., Ansari, M., Akl, E.A., et al. (2019). A risk of bias instrument for non-randomized studies of exposures: A users' guide to its application in the context of GRADE. Environ Int *122*, 168–184. 10.1016/j.envint.2018.11.004.

33. Morgan, R.L., Thayer, K.A., Bero, L., Bruce, N., Falck-Ytter, Y., Ghersi, D., Guyatt, G., Hooijmans, C., Langendam, M., Mandrioli, D., et al. (2016). GRADE: Assessing the quality of evidence in environmental and occupational health. Environ Int *92–93*, 611–616. 10.1016/j.envint.2016.01.004.

34. Committee for Medicinal Products for Human Use (2017). Guideline on strategies to identify and mitigate risks for first-in-human and early clinical trials with investigational medicinal products.

35. Schüssler-Lenz, M., Beuneu, C., Menezes-Ferreira, M., Jekerle, V., Bartunek, J., Chamuleau, S., Celis, P., Doevendans, P., O'Donovan, M., Hill, J., et al. (2016). Cell-based therapies for cardiac repair: a meeting report on scientific observations and European regulatory viewpoints. Eur J Heart Fail *18*, 133–141. 10.1002/ejhf.422.

36. Center for Biologics Evaluation and (2020). Considerations for the Design of Early-Phase Clinical Trials of Cellular and Gene Therapy Products. U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-design-early-phase-clinical-trials-cellular-and-gene-therapy-products.

37. Emmerich, C.H., Gamboa, L.M., Hofmann, M.C.J., Bonin-Andresen, M., Arbach, O., Schendel, P., Gerlach, B., Hempel, K., Bespalov, A., Dirnagl, U., et al. (2021). Improving target assessment in biomedical research: the GOT-IT recommendations. Nat Rev Drug Discov *20*, 64–81. 10.1038/s41573-020-0087-3.

38. van Gerven, J., and Cohen, A. (2018). Integrating data from the Investigational Medicinal Product Dossier/investigator's brochure. A new tool for translational integration of preclinical effects. Br J Clin Pharmacol *84*, 1457–1466. 10.1111/bcp.13529.

39. Goodman, S.N., and Gerson, J. (2013). Mechanistic Evidence in Evidence-Based Medicine: A Conceptual Framework (Agency for Healthcare Research and Quality (US)).

40. Parkkinen, V.-P., Wallmann, C., Wilde, M., Clarke, B., Illari, P., Kelly, M.P., Norell, C., Russo, F., Shaw, B., and Williamson, J. (2018). Evaluating Evidence of Mechanisms in Medicine: Principles and Procedures (Springer).

41. Smith, G.C.S., and Pell, J.P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ *327*, 1459–1461.

42. London, A.J., and Kimmelman, J. (2015). Why clinical translation cannot succeed without failure. eLife *4*, e12844. 10.7554/eLife.12844.

43. Kimmelman, J. (2009). Gene Transfer and the Ethics of First-in-Human Research: Lost in Translation (Cambridge University Press) 10.1017/CBO9780511642364.

44. Dirnagl, U., and Endres, M. (2014). Found in Translation. Stroke *45*, 1510–1518. 10.1161/STROKEAHA.113.004075.

45. Henderson, V.C., Kimmelman, J., Fergusson, D., Grimshaw, J.M., and Hackam, D.G. (2013). Threats to Validity in the Design and Conduct of Preclinical Efficacy Studies: A Systematic Review of Guidelines for In Vivo Animal Experiments. PLOS Medicine *10*, e1001489. 10.1371/journal.pmed.1001489.

46. Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference (Houghton, Mifflin and Company).

47. Parkkinen, V.-P., and Williamson, J. (2020). Extrapolating from Model Organisms in Pharmacology. In Uncertainty in Pharmacology: Epistemology, Methods, and Decisions Boston Studies in the Philosophy and History of Science., A. LaCaze and B. Osimani, eds. (Springer International Publishing), pp. 59–78. 10.1007/978-3-030-29179-2_3.

48. Hooijmans, C.R., Rovers, M.M., de Vries, R.B., Leenaars, M., Ritskes-Hoitinga, M., and Langendam, M.W. (2014). SYRCLE's risk of bias tool for animal studies. BMC Medical Research Methodology *14*, 43. 10.1186/1471-2288-14-43.

49. OHAT Risk of Bias Rating Tool for Human and Animal Studies National Toxicology Program. https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/riskbias.

50. Findley, M.G., Kikuta, K., and Denly, M. (2021). External Validity. Annual Review of Political Science *24*, 365–393. 10.1146/annurev-polisci-041719-102556.

51. Armstrong, J.S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. International Journal of Forecasting *22*, 583–598. 10.1016/j.ijforecast.2006.04.006.