

## Assessing Risk/Benefit for Trials Using Preclinical Evidence: A Proposal

Jonathan Kimmelman\* and Valerie C. Henderson

### Abstract

Moral evaluation of risk/benefit in early phase studies requires assessing the clinical promise of a candidate intervention using preclinical evidence. Yet there is little to guide ethics committees, investigators, sponsors or other stakeholders morally charged with making these assessments (“evaluators”). In what follows, we draw on published guidelines for preclinical study design to develop a structured process for assessing the clinical promise of new interventions. In the first step, evaluators gather all relevant preclinical studies, assess the magnitude of treatment effects, and determine clinical promise in light of various threats to valid clinical inference. In the second step, evaluators adjust assessments of clinical promise from preclinical studies by examining how other agents in the same reference class—and supported by similar evidence—have fared in clinical development. Assessments of clinical promise can then be fed into moral evaluation of risk and benefit in early phase trials. Though our approach has limitations, it offers a systematic and transparent method for assessing risk/benefit in early phase trials of novel interventions.

### Keywords

Ethics, research; Drug evaluation, preclinical; Reproducibility of results; Clinical trial, phase 1; Clinical trial, phase 2; Risk

### Introduction

All policies in trial ethics require a favorable balance of risk and benefit. In many research arenas, judgments about risk/benefit are supported by prior trials and clinical experience. In novel arenas, however, investigators, ethics review committees, sponsors and data monitoring committees (“evaluators”) must draw on preclinical evidence. Nevertheless, prescription on how evaluators should apply preclinical evidence toward risk/benefit assessment is limited.

What follows centers on how evaluators should use evidence of clinical promise gathered in animals (“efficacy studies”). We describe when efficacy studies are crucial for ethical evaluation of trials, and argue that evaluators should directly assess efficacy evidence when

---

\* corresponding author: Studies for Translation, Research, Ethics, and Medicine (STREAM) Group, Biomedical Ethics Unit/Social Studies of Medicine, McGill University, 3647 Peel St., Montréal, QC, H3A 1X1, Canada, ph: 1-514-398-3306; mobile: 1-514-953-3306; fax: 1-514-398-8349, jonathan.kimmelman@mcgill.ca; Preference for communication: By email.

**Interest Disclosure:** The authors have no relevant competing interests to declare.

contemplating early phase trials. We then propose a structured process for assessing clinical promise using efficacy studies. We close by describing limitations of our approach.

## Efficacy Studies and Risk/Benefit

Two classes of preclinical evidence generally support clinical development. The first includes toxicology and pharmacokinetics. These aim at anticipating safety issues and tolerable doses for trials. The second category is “efficacy studies.” These are designed to demonstrate “clinical promise” in a proxy disease setting (we use “clinical promise” in a narrow sense: the probability that an intervention will demonstrate safety and efficacy in human beings after a small number of trials). Generally, the most clinically informative studies are performed in live animals that most closely model human disease. Most of what follows pertains to “efficacy studies.”

Many ethics policies [1][2] urge preceding human investigations with animal studies. These recommendations stem from a recognition that animal studies ground judgments about risk/benefit. First, they are designed to predict clinical effects, and exposing patients to an unproven and possibly noxious intervention is only defensible where there is justified belief in its promise. Second, they clarify the appropriate application of a drug, like dose or indication. Third, they support interpretation of trials; if drugs prove ineffective, researchers can use efficacy studies to determine why.[3]

Despite a patent relationship between efficacy studies and human protections, , there is little to guide evaluators on assessment. Some might hold that evaluation should fall solely to drug regulators, rather than investigators, sponsors or ethics committees. However, drug regulators have detailed requirements for toxicology studies, but not for most efficacy studies. The FDA’s guidance on Investigational New Drugs submissions (INDs) states that “lack of... potential effectiveness information should not generally be a reason for a Phase 1 IND to be placed on clinical hold”[4]. FDA does have more demanding requirements for cell and gene therapies[5], as well as for products licensed for human use without trials.[6] Of course, FDA can exercise discretion in the preclinical evidence it demands through, for example, pre-IND meetings with sponsors. Second, while FDA is tasked with assessing risk, its mandate does not extend to weighing risk against knowledge value. Drug regulators lack authority- and likely the resources and expertise- to balance scientific utilities of a trial against burdens; instead this responsibility is delegated to others.[7] The FDA Guidance for IRBs and Clinical Investigators states “21 CFR 56.111(a)(2) requires the IRB to assure that the risks to subjects are reasonable in relation to the anticipated benefits. The risks cannot be adequately evaluated without review of the results of previous animal and human studies.” Many investigators and ethics committees may currently lack the expertise to review preclinical studies; we address this concern below- for now noting that sponsors, investigators, and others also bear moral responsibility for risk/benefit and thus constitute audiences for the structured evaluation process we propose.

Under existing policy, evaluators should weigh risks to subjects against benefits to them (if any) and to society. Both direct benefits as well as benefits to society in early phase trials are related to the clinical promise of an intervention. Thus, studies involving greater risk

demand a greater belief in an intervention's clinical promise; studies that deliver very small doses (e.g. microdoses) can demand less evidence of promise. At the opposite extreme are early phase studies employing active and prolonged exposures or involving invasive delivery. Even in the presence of clinical need, very strong evidence of efficacy is needed to justify trial risks. Somewhere between are studies where there is limited clinical evidence of promise. Consider drugs that have reached phase 2. Phase 1 studies will have supplied limited evidence of safety and perhaps pharmacodynamic, evidence of activity. However, phase 2 studies deliver unproven drugs to larger cohorts of patients-often at doses close to the limit of tolerability; often they are launched in medical indications that have not been tested in phase 1.[8] Despite having advanced into middle phases of testing, moral justification of risk continues to require an assessment of animal studies.

## Reviewing Efficacy Studies

Though review of efficacy studies is technically demanding, there are widely shared and cardinal considerations that evaluators can use. These derive from an understanding of what efficacy studies set out to demonstrate, and ways that they can mislead.

We propose a two-step, evidence-based process whereby reviewers might use preclinical efficacy studies to support inferences about clinical promise. In the first step, reviewers should use efficacy studies to estimate the clinical promise of a new drug. In the second step, reviewers should adjust their estimates based on clinical outcomes with other, related drugs pursued on similar evidence (reference class outcomes). All else being equal, the stronger the clinical promise, the stronger the moral justification for riskier trials. Estimates of clinical promise can then be joined with other moral judgments in evaluating risk and benefit.

### Step 1: Evaluating Clinical Promise Using Efficacy Evidence

Preclinical efficacy studies set out to provide evidence of clinical promise. They do this by supporting inferences that a candidate intervention causes treatment effects in a model setting, and that these cause and effect relationships generalize to clinical settings. All else being equal, interventions that cause larger effects in animals have greater clinical promise. However, inferences about clinical promise can be threatened in three different ways, and evaluation of preclinical studies begins by determining the extent to which these threats have been addressed during study design and execution.

First, preclinical efficacy studies may introduce biases or random error that lead to spurious causal inferences- these are called "threats to internal validity." For instance, unless blinded to treatment allocation in an experiment, investigators committed to a hypothesis can unknowingly attend more closely to therapeutic responses in treated animals. Second, efficacy studies can fail clinical generalization if investigators mischaracterize the relationship between a preclinical study and an ensuing trial. These mischaracterizations are called "construct validity threats." A researcher might mischaracterize a drug's utility for chronic human disease if it was tested only in animals during acute illness. Third, efficacy studies can fail clinical generalization because there is something idiosyncratic about test conditions that fails to carry over to trials. These artefacts represent "threats to external

validity.” A drug might fail to generalize because of differences in animal physiology or laboratory practices that interact with treatment effects.

The first task in assessing clinical promise begins with evaluators thoroughly retrieving all available and relevant efficacy studies on an intervention (though investigators often reference preclinical evidence in trial brochures, this information can be abbreviated and selective). Evaluators should then determine the magnitude of treatment effects, and interrogate preclinical studies to determine the extent to which preclinical efficacy studies have addressed all three threats to valid clinical generalization. We recently performed a systematic review of guidelines for designing clinically generalizable efficacy studies, and identified 26 guidelines pertaining to 11 broad disease areas, such as neurological and cerebrovascular diseases or cardiac and circulatory diseases. We also produced a minimal set of 14 consensus recommendations for the three types of validity threat.[9] If a trial involves a disease area addressed by a focused guideline, evaluators can evaluate the clinical promise for a given intervention based on whether methods in the guideline(s) have been implemented. Alternatively or in combination with specific guidelines, reviewers can tap the collective wisdom of the preclinical research community and draw on our 14 consensus recommendations.

In particular, reviewers should assess the extent to which efficacy studies have minimized threats to internal validity by asking at least the following six questions: 1) was the sample size sufficient to exclude random variation as an explanation of effect sizes? 2) were animals randomly assigned to treatment? 3) were outcome assessors blinded? 4) do studies explain the flow of animals from inclusion through to analysis? 5) were appropriate controls included? and 6) was a dose-response relationship tested?

Reviewers can address the extent to which construct validity threats were addressed by at least examining whether, 1) animals were characterized at baseline and included based on pre-specified eligibility criteria, 2) animal models used are widely believed to offer the most faithful and available representation of the human disease that will be tested in trials, 3) studies establish a mechanism of action, 4) outcomes used in studies are good representations of clinical outcomes used in trials, and 5) animal age is matched to age of patients (often, it is possible to scale an animal’s age to that of a human being, based on developmental milestones, longevity, etc.; see, for instance [10]).

The extent to which external validity threats have been addressed can be assessed by determining whether efficacy studies were replicated in 1) more than one model, 2) more than one laboratory, and 3) more than one species. The importance of testing in more than one system is underscored by the fact that animal models are often flawed representations of human disease and may capture a narrow set of disease phenomena. Observing activity in multiple models provides greater confidence that activities are robust and address more fundamental disease processes.

Failure to implement any one of the above practices in the context of otherwise promising effect sizes should qualify a reviewer’s confidence that activities in animal studies will generalize to patients.

## Step 2: Combining with Reference Class Evidence

At this point, a reviewer will have formed an opinion about a drug's clinical promise using efficacy evidence alone. However, drug candidates backed by strong animal evidence frequently fail human trials; conversely, drug candidates backed by weak animal evidence may succeed in well understood realms (e.g. "me too drugs"). Often, this is because animal models or the circumstances of their use are poor representations of the clinical scenarios. The second step is to adjust assessments of clinical promise from efficacy studies by examining how other agents in the same class- and supported by similar evidence- have fared in clinical development. If most drugs in the same reference class have been successfully translated on similar evidence (a sign that animal models used are predictive), this provides grounds for confidence that responses in animals will be recapitulated in human beings. If most drugs in the same reference class have not demonstrated efficacy in trials (a sign that animal models used are not predictive), this provides grounds for caution in projecting clinical responses. For instance, imagine researchers are pursuing a cancer drug that inhibits a particular receptor. If several drugs targeting the same receptor were supported by similar preclinical evidence but failed translation, reviewers should be more guarded in their estimation of clinical promise. A discussion of this step- and its relationship with models of disease- is described elsewhere.[11]

The next task is to feed estimates of clinical promise into evaluations of risk/benefit. A thorough account of this task is well beyond the scope of this paper and has been discussed by other thoughtful commentators, including in early phase clinical development. [12–14] We nevertheless offer four broad factors that should inform the moral evaluation of risk/benefit. The first is the moral justification for risks of an intervention. If an intervention implicates care obligations (e.g. a intervention will be substituted for an established effective therapy), then the ethical permissibility of its risks is to be judged on whether the clinical promise is competitive with standard of care. In previous work, one of us has argued that the moral justification for exposing patients to novel interventions in early phase studies will generally rest on the value of the knowledge a study is expected to produce, not therapeutic value.[15] This then prompts a second consideration: risk to subjects, which is related to subjects' medical options and expected research burdens. A third consideration is the degree to which a research program opens up new avenues for investigation. For highly novel intervention platforms or targets, trials can often turn up findings that are applicable across other translation efforts even if they fail to recapitulate preclinical studies. A last factor is the expected value to healthcare systems should the intervention translate successfully. Studies involving interventions that are believed to have transformative potential- or that address priority unmet health needs- permit greater subject risk than those pursuing incremental innovations.[13, 14] Our proposal for evaluating risk/benefit using efficacy studies is depicted in Figure 1.

## Implementation

We offer a hypothetical trial proposal to briefly illustrate how our approach might be applied. A researcher proposes a phase 1 trial investigating a novel cell therapy strategy for treatment of myocardial infarction (MI). A systematic retrieval of preclinical studies turns up

one that showed large responses in a “gold standard” swine model of MI. The swine study was randomized and blinded, demonstrated the mechanism of action, and used appropriate endpoints. However, the study used a small sample and has not been replicated, and it used a surrogate endpoint-reduction in infarct size- rather than a clinically relevant measure of efficacy. The evaluator concludes her review of preclinical evidence with a somewhat qualified level of confidence in clinical promise. Next, the evaluator sets this confidence against past performance of cell therapies in MI, which have generally shown safety but clinically marginal benefit.[16] The evaluator adjusts her confidence in clinical promise still downwards. She now feeds this into her evaluation of risk/benefit, which considers the expected burdens of the trial, the magnitude of unmet need for MI treatments, the medical options available for eligible patients, and the overall scientific fecundity of this line of investigation. An unfavorable judgment might lead the evaluator to condition acceptance on the completion of more rigorous and favourable preclinical studies, or to employ a less aggressive research strategy (e.g. starting at lower doses, or enrolling patients with more advanced disease). If the evaluator is a clinical investigator, it might lead her to refuse to join the protocol and enroll patients.

## Discussion

Our proposal offers a structured approach to using efficacy studies when assessing risk/benefit for trials. By combining all accessible efficacy studies with reference class information, it capitalizes on the totality of evidence. In these respects, it offers a systematic and relatively transparent method for incorporating preclinical findings into the moral evaluation of trial risk. It also removes some of the arbitrariness that likely plagues a process for which there are no established frameworks.

It might be objected that ethics committees lack the expertise to perform these evaluations. We stress that our analysis is intended to apply to any stakeholder who bears moral responsibility for maintaining a favourable risk/benefit balance in trials- including public funders, private sponsors and trial steering committees- who surely have the expertise. Insofar as ethics committees have a role in refereeing such judgments, we suggest that the process of review can be facilitated by presenting study sponsors with a short template asking them to address each element listed in appendix 1; such templates are similar to those used in submission of manuscripts at biomedical journals like *Nature*.<sup>[17]</sup>

Such assessments might also be delegated to others.. For instance, institutions overseeing early phase trials might establish bodies that provide “special scrutiny;” they might include nonvoting ad hoc expert reviewers in deliberations<sup>[18]</sup>. or they could require that sponsors proposing early phase protocols submit an expert and independent scientific review of the preclinical evidence. The prescription to maintain a favorable risk/benefit balance in trials should not rest solely on the shoulders of ethics committees. Our analysis is not aimed at defining appropriate institutional mechanisms; in any event, these will vary depending on contextual factors of review.

The approach has important limitations and demands further refinement, of course. First, we addressed only the first step in risk/benefit evaluation- namely, assessment of clinical



promise. As already described, this judgment must be combined with other judgments to morally evaluate the risk/benefit. Second, the approach should not be applied mechanically. Experimental practices in step 1 do not exhaust all factors that threaten strength of evidence; also, some practices matter more in some settings than others (e.g. nonrandomization is a greater liability in studies involving outbred animals). Third, reviewers should be alert to the fact that preclinical research is afflicted with publication and reporting bias,[19] and some efficacy studies may be impossible to access. Given publication biases in preclinical research, evaluators should approach investigators and/or sponsors for preclinical efficacy evidence that might not be reflected in the study brochure. Fourth, the approach has not been rigorously validated. Considerations enumerated in step 1 are derived from a systematic review- and many of these recommendations are supported by evidence.[20] Nevertheless, there are established methods for validating guidelines, and the proposal above should be put to such a process. Fifth, our approach assumes that consensus recommendations contained in preclinical guidelines have identified the most pressing validity threats. It is possible these guidelines overlook larger or more pressing validity threats. Last, our approach to reviewing preclinical studies is embedded within a potentially flawed, standard research ethics framework. In this framework, review committees and investigators are not instructed to contemplate opportunity costs or externalized costs of investigation. Nor are they asked to contemplate the knowledge value of a protocol detached from subject risk.[21] As a consequence, the approach we have described could sanction resource intensive but marginally informative early phase studies- provided they do not expose subjects to undue risk.

Investigators and reviewers are charged with promoting a favorable balance of risk/benefit in trials, and this process begins with systematic assessment of clinical promise. Nowhere is this task more challenging than in early phase trials. Fortunately, preclinical researchers themselves have articulated criteria for determining strength of evidence of clinical promise. Our approach absorbs this collective wisdom into a structured method for assessing the probabilistic component of risk/benefit in early phase trials.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

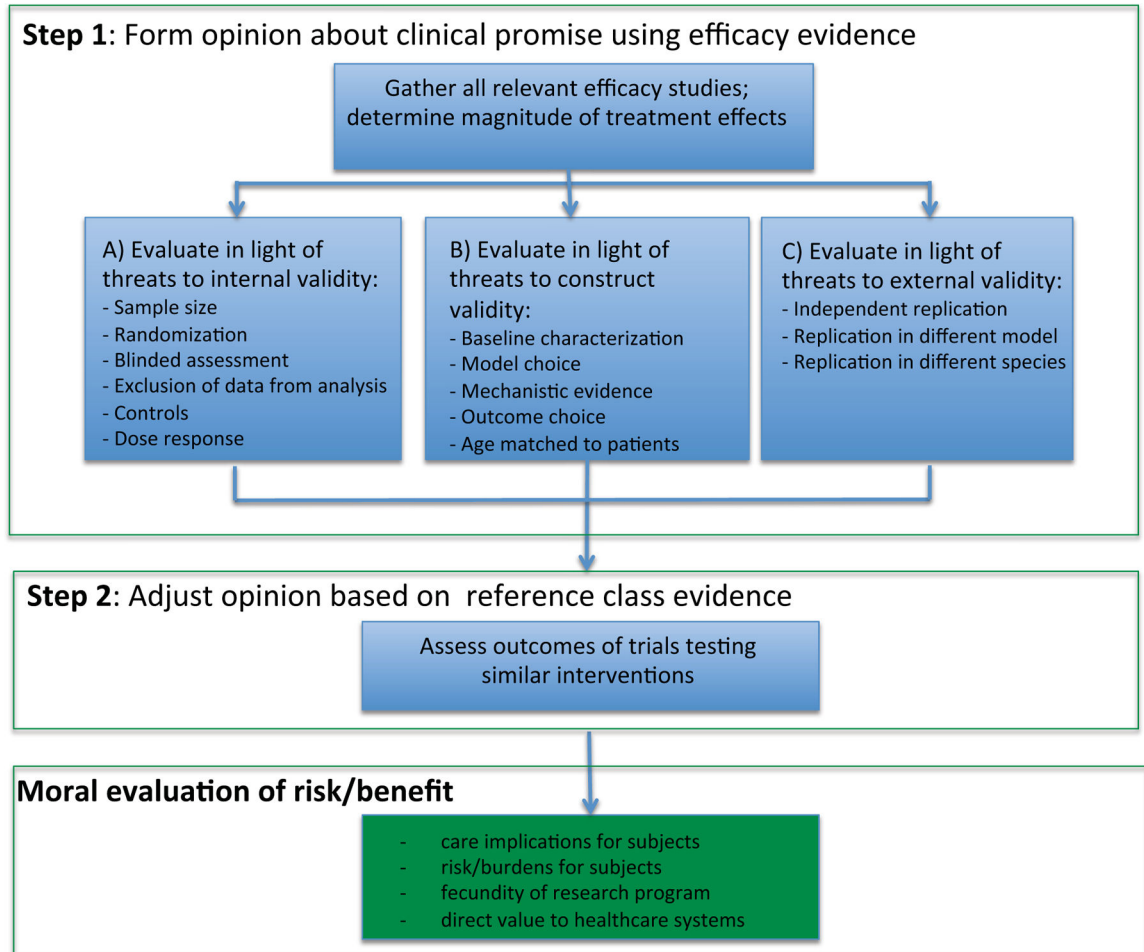
**Funding:** This work was funded by the Canadian Institutes of Health Research (EOG 111391). The funder had no role in the preparation of the manuscript or the decision to publish. We thank Chris McCabe for comments on an earlier draft; faults that remain are our own.

## References

1. World Medical Association. Declaration of Helsinki. *JAMA : the journal of the American Medical Association*. 2013; 310(20):2191–4. published Online First: Epub Date. DOI: 10.1001/jama.2013.281053 [PubMed: 24141714]
2. Trials of war criminals before the Nuremberg military tribunals under control council law. Washington, D.C: U.S. Government Printing Office; 1949.
3. Kimmelman J, Anderson JA. Should preclinical studies be registered? *Nature biotechnology*. 2012; 30(6):488–9. published Online First: Epub Date. DOI: 10.1038/nbt.2261

4. Services USDoHH. Guidance for Industry: Content and Format of Investigational New Drug Applications (INDs) for Phase 1 Studies of Drugs, Including Well-Characterized, Therapeutic, Biotechnology-derived Products. 1995.
5. Services USDoHH. Guidance for Industry: Preclinical Assessment of Investigational Cellular and Gene Therapy Products. 2013.
6. U.S. Department of Health and Human Services FaDA, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for Industry: Product Development Under the Animal Rule. 2014.
7. Cavazzana-Calvo M, Fischer A. Gene therapy for severe combined immunodeficiency: are we there yet? *The Journal of clinical investigation*. 2007; 117(6):1456–65. [PubMed: 17549248]
8. Ratain MJ, Mick R, Schilsky RL, et al. Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. *Journal of the National Cancer Institute*. 1993; 85(20):1637–43. [PubMed: 8411243]
9. Henderson VC, Kimmelman J, Fergusson D, et al. Threats to validity in the design and conduct of preclinical efficacy studies. *PLoS medicine*. 2013; 10(7):e1001489. published Online First: Epub Date. doi: 10.1371/journal.pmed.1001489 [PubMed: 23935460]
10. Sengupta P. The Laboratory Rat: Relating Its Age With Human's. *International journal of preventive medicine*. 2013; 4(6):624–30. [PubMed: 23930179]
11. Kimmelman J, London AJ. Predicting harms and benefits in translational trials. *PLoS medicine*. 2011; 8(3):e1001010. published Online First: Epub Date. doi: 10.1371/journal.pmed.1001010 [PubMed: 21423344]
12. Dresser R. Building an ethical foundation for first-in-human nanotrials. *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics*. 2012; 40(4): 802–8. published Online First: Epub Date. DOI: 10.1111/j.1748-720X.2012.00708.x
13. Habets MG, van Delden JJ, Bredenoord AL. The social value of clinical research. *BMC medical ethics*. 2014; 15:66. published Online First: Epub Date. doi: 10.1186/1472-6939-15-66 [PubMed: 25189994]
14. Casarett DJ, Karlawish JH, Moreno JD. A taxonomy of value in clinical research. *Irb*. 2002; 24(6): 1–6.
15. Anderson JA, Kimmelman J. Extending clinical equipoise to phase 1 trials involving patients: unresolved problems. *Kennedy Inst Ethics J*. 2010; 20(1):75–98. [PubMed: 20506695]
16. Clifford DM, Fisher SA, Brunskill SJ, et al. Stem cell treatment for acute myocardial infarction. *The Cochrane database of systematic reviews*. 2012; 2:CD006536. published Online First: Epub Date. doi: 10.1002/14651858.CD006536.pub3
17. For Authors Manuscript Formatting Guide. Secondary For Authors Manuscript Formatting Guide. <http://www.nature.com/nature/authors/gta/-a5.9>
18. Levine C, Faden R, Grady C, et al. “Special scrutiny”: a targeted form of research protocol review. *Annals of internal medicine*. 2004; 140(3):220–3. [PubMed: 14757620]
19. Sena ES, van der Worp HB, Bath PM, et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *Plos Biol*. 2010; 8(3):e1000344. published Online First: Epub Date. doi: 10.1371/journal.pbio.1000344 [PubMed: 20361022]
20. Crossley NA, Sena E, Goehler J, et al. Empirical evidence of bias in the design of experimental stroke studies. *Stroke; a journal of cerebral circulation*. 2008; 39(3):929–34. published Online First: Epub Date. DOI: 10.1161/STROKEAHA.107.498725
21. London AJ, Kimmelman J, Carlisle B. Research ethics. Rethinking research ethics: the case of postmarketing trials. *Science*. 2012; 336(6081):544–5. published Online First: Epub Date. DOI: 10.1126/science.1216086 [PubMed: 22556237]





**Figure 1.** Process for assessing clinical promise and risk/benefit during trial evaluation.