The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration

Kent, D.M.; Klaveren, D. van; Paulus, J.K.; D'Agostino, R.; Goodman, S.; Hayward, R.; ... ; Steyerberg, E.W.

Published in final edited form as:
Ann Intern Med. 2020 January 07; 172(1): W1–W25. doi:10.7326/M18-3668.

The PATH Statement Explanation and Elaboration Document

David M. Kent, MD, MS[1], **David van Klaveren, PhD**[1,12], **Jessica K. Paulus, ScD**[1], **Ralph D'Agostino, PhD**[2], **Steve Goodman, MD, MHS, PhD**[3], **Rodney Hayward, MD**[4], **John P.A. Ioannidis, MD, DSc**[3], **Bray Patrick-Lake, MFS**[5], **Sally Morton, PhD**[6], **Michael Pencina, PhD**[5], **Gowri Raman, MBBS, MS**[7], **Joseph S. Ross, MD, MHS**[8], **Harry P. Selker, MD, MSPH**[9,10], **Ravi Varadhan, PhD**[11], **Andrew Vickers, PhD**[13], **John B. Wong, MD**[14], **Ewout W. Steyerberg, PhD**[12]

[1]Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center, Boston, MA, USA
[2]Department of Biostatistics and Epidemiology, Boston University, Boston, MA, USA [3]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[4]Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA
[5]Duke Clinical Research Institute (DCRI), Duke University, Durham, NC, USA [6]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA [7]Center for Clinical Evidence Synthesis, ICRHPS, Tufts Medical Center, Boston, MA, USA [8]Section of General Medicine, Department of Internal Medicine, School of Medicine and Department of Health Policy and Management, School of Public Health, Yale University, New Haven, CT, USA
[9]Center for Cardiovascular Health Services Research, ICRHPS, Tufts Medical Center, Boston, MA, USA [10]Tufts Clinical and Translational Science Institute, Boston, MA, USA [11]Center on Aging and Health, Johns Hopkins University, Baltimore, MD, USA [12]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands [13]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[14]Division of Clinical Decision Making, Tufts Medical Center, Boston, MA, USA

Abstract

The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement was developed to promote the conduct of, and provide guidance for, predictive analyses of heterogeneous treatment effects (HTE) in clinical trials. The goal of predictive HTE analysis is to provide patient-centered estimates of outcome risks with versus without the intervention, taking into account all relevant patient attributes simultaneously, to support more personalized clinical decision making than can be made based only on an overall average treatment effect. We distinguished two categories of predictive HTE approaches (a "risk modeling" and an "effect modeling" approach) and developed four sets of guidance statements: 1) criteria to determine when risk modeling approaches are likely to identify clinically meaningful HTE; 2) methodological aspects of risk modeling methods; 3) considerations for translation to clinical practice; and 4) considerations and caveats in the use of effect modeling approaches. We also discuss limitations of these methods and enumerate research

Corresponding Author: David M Kent, MD, MS, Director, Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, 800 Washington St, Box 63, Boston, MA 02111, dkent1@tuftsmedicalcenter.org.

priorities for advancing methods designed to generate more personalized evidence. This explanation and elaboration document describes the intent and rationale of each recommendation; and discusses related analytic considerations, caveats, and reservations.

Keywords

personalized medicine; subgroup analysis; RCTs; heterogeneity of treatment effect; predictive analytics

Introduction

In medical care, treatment decisions made by clinicians and patients are generally based implicitly or explicitly on predictions of comparative outcome risks under alternative treatment conditions. Randomized controlled trials (RCTs), widely accepted as the gold standard for determining causal effects, have provided the primary evidence for these predictions. However, there has been mounting recognition within Evidence-Based Medicine (EBM) of the limitations of RCTs as tools to guide clinical decision making at the individual patient level.[1–4] Although historically the overall summary result from randomized trials ("average treatment effect") has been the cornerstone of evidence-based clinical decisions, there is growing interest in understanding how a treatment's effect can vary across patients —a concept described as heterogeneity of treatment effect (HTE).[5–11]

There is a large literature on the limitations of conventional "one-variable-at-a-time" subgroup analyses, which serially divide the trial population into groups (e.g., male versus female; old versus young) and examine the contrast in the treatment effect across these groups.[12–22] The limitations include the risks of false negative and false positive results— due to low power for statistical interactions, weak prior theory on potential effect modifiers and multiplicity.[4;10;23–25] These analyses also are incongruent with the way clinical decision making occurs at the level of the individual patient, since patients have multiple attributes simultaneously that can affect the trade-offs between benefits and harms of the intervention. Individual patients therefore may belong to multiple different subgroups, each potentially yielding different treatment effect estimates.[4;10]

The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement offers guidance relevant for "predictive" approaches to HTE analysis[26] that are designed to address some of the limitations mentioned above. The goal of predictive HTE analysis is to provide individualized predictions of treatment effect, specifically defined by the difference between expected potential outcome(s) of interest with one intervention versus an alternative.[4;8] We refer to this as the "individualized treatment effect" and avoid the term "individual treatment effects," since this latter term confusingly suggests that treatment effects can be estimated at the person-level; such effects are inherently unobservable in parallel arms clinical trials.[10;27] Individualized treatment effects have also been termed "conditional average treatment effects" [28]—denoting that they are the averaged treatment effect in a subpopulation (i.e. conditioned on a set of covariates). However, for prediction, we are specifically interested in identifying the best conditional average treatment effect given all available patient characteristics, where 'best' is defined as that which best discriminates between *future* patients who do and do not benefit from a treatment to optimize individual patient decision making. By accounting for multiple variables simultaneously, predictive HTE analysis is foundational to the concept of personalization in EBM.[4]

Distinct approaches to PATH

The PATH Statement outlines a set of principles, criteria, and key considerations for predictive approaches to HTE in RCTs to provide patient-centered evidence in support of decision making. The focus of the PATH Statement guidance is on identifying "clinically important HTE,"[4;7;10] or variation in the risk difference across patient subgroups sufficient to span important decision thresholds, which reflect treatment-related harms and burdens. The PATH Statement[26] offers guidance on two distinct approaches to predictive HTE analysis.[4] With a "risk modeling" approach, first, a multivariable model that predicts the risk of an outcome (usually the primary study outcome) is identified from external sources (an "external model") or developed directly on the trial population without a term for treatment assignment (an "internal model"). This prediction model is then applied to disaggregate patients within trials to examine risk-based variation in treatment effects. In a second, "effect modeling," approach, a model is developed on RCT data with inclusion of a treatment assignment variable, and potential inclusion of treatment interaction terms. These more flexible effect modeling approaches have the potential to improve discrimination of patients who do and do not benefit but are especially vulnerable to overfitting and to false discovery of promising subgroup effects (or require very large databases well powered for the detection of interaction effects)[29]. Both approaches can be used to predict individualized treatment effects, i.e., the *difference* in expected outcome risks under two alternative treatments, conditional on multiple important clinical variables. A fuller introduction to risk and effect modeling is presented in prior literature.[4]

In this PATH Statement Elaboration and Explanation, we expand on the intent and motivation (and reservations) regarding the PATH statements, criteria, and considerations and caveats. Recommendations are explained in more detail and accompanied by clinical applications of selected methods and supporting methodological evidence where relevant. A glossary of terms relevant to these methods is also provided in an Appendix.

Clarification of Terms and PATH Statement Scope

The term heterogeneous treatment effects (HTE) has been used in the literature in different ways. In this paper, we define the term HTE as non-random variation of treatment effects across levels of a covariate (i.e., a patient attribute or a score comprised of multiple attributes), as measured on a selected scale, against a clinical outcome. It corresponds to the epidemiological concept of effect measure modification, but applies specifically to treatment effects. HTE is identified in clinical trials by contrasting treatment effects on a chosen scale between subgroups and testing for statistical interactions. Importantly, HTE, effect measure modification, and statistical interaction are all 'scale-dependent' concepts—that is, their presence or absence depends on what scale one selects to measure treatment effect.[30] The scale dependence of HTE is illustrated in Figure 1, which contrast three different analyses (the first showing HTE only on the absolute scale, the second only on the relative scale, and the third on both). To underscore the scale dependence of HTE, we also show the results of a risk modeling analysis of the Diabetes Prevention Program (DPP) Trial, which tested lifestyle modification and Metformin against usual care for prevention of diabetes (Figure 2). While only one of the tested therapies shows statistically significant HTE on the relative (hazard ratio) scale, both therapies had substantial HTE on the clinically important absolute scale. The Appendix Glossary contains definitions of key terms described here and elsewhere in the PATH Statement and E&E document.

HTE analysis for causal interaction versus for prediction and decision making

We also note that HTE (and statistical interactions) are used to make two very different kinds of inferences: 1) causal inferences (e.g. regarding causal/biological interaction) and 2) inferences for clinical decision making. Although the importance of statistical interactions is often stressed for HTE analysis, we note that these inferences are only weakly related to the presence of "statistically significant HTE." Statistical interactions should not be confused with causal interactions and statistically significant HTE should also not be conflated with clinically-important HTE. These issues are described briefly below.

In regression models examining HTE, **causal inferences** depend on interpretation of model *inputs* (i.e. model covariates). *The PATH guidance does not address causal interpretations of HTE*. These analyses are important for identifying biomarkers that might biologically interact with therapy. Interaction on a multiplicative (relative) scale is taken by many as being stronger evidence in support of a causal interaction than interaction on an absolute scale (although this is by no means a universal view).[31–37] Nevertheless, we note that treatment-by-covariate interactions (on any scale) are generally *descriptive* measures of association (when the covariate is not randomly assigned, as in a factorial trial), since an interacting covariate may be acting as a proxy for many measured and unmeasured variables. To attribute a change in the treatment effect to the covariate, one would need to control for all differences in these other variables (i.e. observed and unobserved confounders) across levels of the subgrouping factor. In any event, demonstrating causal interaction is not necessary for "predictive" HTE analyses that seek to target therapies to those who most benefit (see immediately below).

In regression models examining HTE, **inferences for clinical decision making** depend on interpretation of model *outputs*. Because these analyses depend on model outputs, they have been referred to as "predictive" HTE analyses.[4;8] *The PATH guidance is limited to predictive approaches to HTE.* The goal of predictive HTE analysis is to develop models that can be used to predict which of two or more treatments will be better for a particular individual, taking into account multiple relevant variables.[4;8] Clinically-important HTE occurs when variation in the risk difference across patient subgroups spans a decisionally-important threshold, which depends on treatment burden (including treatment-related harms and costs). It is generally assessed on the absolute scale, regardless of the scale of the analysis. The scale dependence of effect heterogeneity is also illustrated in Figure 1. We also note that controlling for confounding factors (i.e. those factors that differ between levels of the subgrouping variable) is not necessary for prediction.[32;38]

A new term, "risk magnification," has recently been coined to describe a method of identifying high risk-high benefit patients. [39;40] This approach depends on the observation that relative effects (and in particular those on the odds ratio scale) are often more stable than absolute effects.[41–43] Risk magnification is distinct from the risk modeling approach described here, since it may be applied without any data based on the assumption of a contant relative treatment effect. Indeed, the use of the Pooled Cohort Equation[44] (also known as the Atherosclerotic Cardiovascular Disease [ASCVD] Risk Estimator) to target statin therapy to patients at high risk of developing coronary heart disease might be described as an application of "risk magnification." Because many (observed and unobserved) patient attributes change across different risk levels, and because the causes of the outcomes may also change, the assumption of a consistent treatment effect across all levels of risk is a strong assumption that should ideally be examined using randomized data. Additionally, the rate of adverse events may also differ across levels of baseline risk (see Box B**, Recommendation 9**). Examining randomized data stratified by a risk model also permits these other (non-primary) outcomes to be examined across levels of risk.

PATH Statement Criteria for When Risk Modeling is Likely to be of Value (Box 1)

The below criteria identify the data, modeling, and clinical decisional features common to scenarios in which the application of predictive HTE analyses to treatment comparisons is likely to be relevant for individualized clinical decision-making. The motivation and reservations regarding these criteria (and excluded ones) are elaborated.

1. When there is a well-established overall treatment effect. Subgroup results (including risk-based subgroup results) from overall null trials should be interpreted cautiously.

When clinical trials are null, there may be a temptation to find subgroups of patients in whom the treatment might work. However, clinically important subgroup effects discovered through a risk modeling approach are likely to be rare when treatment efficacy has not been established. For example, among 18 null trials in a recent study, risk modeling did not yield clinically informative results on any.[45] More "aggressive" effect modeling approaches (i.e., those reliant on including treatment-by-covariate interaction terms within a prediction model) may identify groups of patients that appear to benefit, but such approaches also are likely to yield spurious false positive results.[28;46] Groups may be suggested by pure chance, aggravated by multiple testing, even when treatments have no effects whatsoever.[29] Thus, predictive HTE analyses are more appropriately conducted on interventions for which an overall effect has been established. Despite a tendency to focus on any positive subgroup in an otherwise null trial, in the absence of strong, *a priori* clinical justification, predictive HTE analyses on null trials are unlikely to lead to reliable clinical evidence.

Possible exceptions to this general rule are interventions with known treatment-related harms that mediate primary outcomes in the treatment arm that might nullify an overall effect. For example, ACE inhibitors both cause and prevent renal insufficiency; thrombolytics both cause (via hemorrhage) and prevent (via reperfusion) functional disability in patients presenting with acute stroke;[47]

carotid surgery can both cause and prevent ischemic stroke in patients with carotid stenosis;[48] and antiarrhythmic agents both cause and prevent serious cardiac arrhythmias.[49;50] In these circumstances, even when trials are null overall for the average treatment effect, risk models may be helpful in disaggregating patients who benefit from those who are harmed, either through application of an outcome risk model or a model identifying those patients at high risk for treatment-related harm (i.e., for "treatment deselection").[47;51;52] Such an approach is possible particularly when the predictors of outcome risk are poorly (or negatively) correlated with predictors of the risk of treatment-related harm. For example, excluding patients at high risk for thrombolytic-related intracranial hemorrhage in stroke (e.g. younger age, lower blood pressure), may uncover benefit in patients at lower risk (ref).[47]

2. When the benefits and harms/burdens of a given intervention are finely balanced (i.e. of similar magnitude on average), increasing the sensitivity of the treatment decision to risk prediction.

Risk models are more likely to be useful when they support a particular "risk-sensitive" decision. This criterion corresponds to the observation that a prediction model or decision rule has maximum expected utility when the decision threshold is at or near the mean population risk.[53;54] The threshold depends on the clinical context and involves weighing the expected benefits against the expected harms or costs of a decision. This assessment may be done with formal decision analysis, but more often it is done informally.

That the value of a risk model is higher for decisions where the threshold is near the population mean can be intuitively understood by considering that the value depends on the proportion of patients for whom the optimal treatment switches from the treatment that is better on average to the alternative given their model-estimated risk, and the benefits that accrue to those switching.[55–57] Beyond measuring model accuracy, various methods have been proposed as a means to evaluate the potential impact that model use might have on a particular decision, including risk stratification tables,[58] relative utility curves,[59] predictiveness curves,[60] and decision curves.[56]

This discussion with respect to risk prediction also fully applies to benefit prediction. When the overall average benefit in a trial is balanced by the average treatment-related harms and costs (i.e. when *net* benefit is near zero), any additional prognostic/predictive information is likely to be especially useful for determining the better therapy for a particular patient.

The relative utility of risk prediction when average risk is near versus far from a threshold is described schematically in Figure 3, which depicts the distribution of expected benefits when a risk model with a C statistic of 0.75 is applied to a population with an average risk of 25%. We consider two different treatment conditions: 1) a treatment with a moderate average treatment effect (a 15% relative risk reduction; 3.8% risk difference) and a slightly favorable average benefit-harm trade-off (compared to a minimally clinically important difference of 3%), and 2) a treatment with a large treatment effect (a 50% relative risk reduction; 12.5% risk difference) and a clearly favorable average benefit-harm tradeoff. All else being equal, the net benefit of risk modeling would be greater when the harm-benefit trade-offs are more finely balanced since risk-stratification would reveal many patients (almost half the trial population in the first schematic example) whose risk-specific optimal treatment differs from the treatment that is best on average. Preliminary evidence from careful simulations or even simple algebraic calculations (see Figure 4A) using plausible assumptions may be important in motivating research and should generally be included in research proposals and protocols.

We acknowledge that the presentation of a single threshold is a simplification, since the threshold is sensitive to patient values and preferences, and because treatment harms and burdens are likely to vary across risk groups. For example, the CHADS Score (and its variants) are used to target anticoagulation to patients with non-valvular atrial fibrillation, but it is also known that higher-CHADS Score patients are also at higher risk for anticoagulation-related hemorrhage.[61] Given the potential (positive or negative) correlation between the benefits and harms, we recommend that harms should be reported in each risk stratum to support strata-specific evaluation of benefit-harm trade-offs (Box B**, recommendation 9**).

3. When treatments are associated with a non-trivial amount of serious harm or burden, increasing the importance of careful patient selection.

This criterion is related to the previous two in that HTE will be most important to decision making in the presence of a *qualitative* interaction, meaning that some patients benefit while others are harmed. Qualitative interactions, by definition, do not arise where treatments are totally innocuous. In the presence of a small amount of treatment-related harm, the harm may be quantitatively negligible among high risk patients, but sufficient to erode much (or all) of the benefit in low risk patients (Figure 4). The importance of risk modeling for HTE in treatments with treatment-related harm has been demonstrated in simulation studies[46;62] and observed empirically for carotid endarterectomy,[48] stroke prevention in non-valvular atrial fibrillation,[63;64] and medical or mechanical reperfusion in ST-elevation myocardial infarction.[65–67] Treatment-related harm may be reflected in the primary outcome, or may be ascertained as a separate outcome (e.g., acute kidney injury, major hemorrhage,[51;68] serious bone fractures).[52;69] Risk modeling may also be appropriate for particularly burdensome interventions (e.g., major lifestyle commitments,[70;71] treatment-related costs).[72;73]

4. When several large, randomized, well-conducted clinical trials of contemporary interventions are available and appropriate for pooling in individual patient meta-analysis

Clinical trials are typically powered to detect an overall average treatment effect in a population, not estimate effects in relevant subgroups. Very large databases are required for effect modeling, in which multiple individual covariate-by-treatment interactions are considered. The sample size required to detect a subgroup effect is four-fold higher than for a main effect even under favorable conditions (e.g., well-balanced subgroups, overall effect and subgroup effect similar in size), but will generally be much higher than that.[23;24] While a risk modeling approach does not depend on the discovery of statistically significant covariate-by-treatment interaction terms, greater statistical power improves the precision of effect estimation across risk strata, thus improving the ability to estimate benefit- harm trade-offs across strata. We also emphasize the need for contemporary trials, as they are more likely to be relevant for contemporary clinical care. In addition, while such analyses have not been consistently successful in the discovery of reliable treatment effect interactions,[74] combining data is likely to lead to substantial increases in risk heterogeneity in the study population (see Criterion 5), increasing the likelihood of uncovering clinically important HTE (i.e. on the risk difference scale).[75;76]

5. When substantial, identifiable heterogeneity of risk in the trial population is anticipated

Risk heterogeneity is dependent on the presence of significant factors to predict outcome risk and differences in the distribution of these factors across the population (i.e. a non-homogeneous population). In the absence of factors that can predict outcome risk, there is no risk heterogeneity. Conversely, risk heterogeneity is highest in the presence of good discrimination (i.e., a high C-statistic). Indeed, risk heterogeneity in a given population is a model-dependent property.[77] Figure 5 shows the empirical relation between the C-statistic and the extreme quartile risk ratio (EQRR, the ratio of outcome rate in the highest risk quartile to the rate in the lowest risk quartile[78]) across 32 publically available trials.[45] Because there is an abundance of clinical prediction models,[7;79;80] the predictability of trial outcomes can generally be evaluated, at least informally, from the literature. Trials with broad inclusion criteria, and thus a broad case mix, are more likely to show greater risk-heterogeneity compared to trials with more narrowly restrictive enrollment criteria. Nevertheless, there seems to be substantial risk heterogeneity even in classic efficacy trials.[45;66-68;71] Because individual patient meta-analyses (i.e., combining trials) have even higher patient-level risk heterogeneity, they are an ideal substrate for these analyses.[75;76]

6. When there is strong preliminary evidence that a prediction model is clinically useful for treatment selection, or when there are models in current use for treatment selection

The vast majority of prediction models developed are not applied in clinical practice.[79] Some fields produce massive numbers of new prediction models without clear purpose except as publishable analytic exercises. Conversely, the development of new models should start from some clinical need. Thoughtful selection of a “risk-sensitive” decision is one of the most crucial steps in developing a useful clinical prediction model.[81] The use of a prediction model in clinical practice may be an implicit marker of such a risk-sensitive decision for which clinicians sense that the balance of the benefits and burdens of a treatment decision vary across the population in a clinically meaningful way. For example, the widespread use of the CHADS score[63;64] (and its variations[82]), the ASCVD score,[44] and chest pain tools[83;84] may be considered a marker of the risk-sensitivity of these decisions. Similarly, the widespread use of certain diagnostic prediction models in emergency departments to rule out rare but serious conditions (e.g. cervical spine fracture,[85] intracranial hemorrhage,[86] or pulmonary embolism[87]) in low risk patients to reduce the harms/burdens of further diagnostic testing is a marker for the risk-sensitivity of this class of decisions. Such consensually-established, implicitly-revealed risk-sensitive decisions remain relatively uncommon. Moreover, randomized data are relatively scarce and risks may change meaningfully over time. Hence, opportunities to re-examine the risk-specific benefits (or validate predictions of benefit) in new trial data are highly valuable.

7. When the clinical variables in the proposed model are routinely available in clinical care.

The advantages of easily and reliably obtainable clinical characteristics as predictors should be obvious. Nevertheless, there are many examples in the literature of models including variables not ordinarily obtained in clinical care. For example, waist-to-hip ratio is a very strong predictor of diabetes and of cardiovascular risk,[88;89] but it is rarely ascertained in routine clinical care. Prostate volume is an important predictor of prostate cancer risk, but can only be obtained by an invasive test and is therefore of quesitonnable value for use in models.[84] By raising the burden of variable ascertainment, the probability that a prediction model will be used—compared to selecting the best treatment on average—is lowered. Again, because of the abundance of published risk models, it is usually possible to ascertain from the literature well established risk predictors prior to the analysis of trial data, even when internal risk models will be used to stratify the trial.

Explication of Excluded Criteria—The PATH Technical Expert Panel (TEP) failed to reach consensus (as defined by a mean agreement score of less than 3) on two additional criteria to identify when a risk modeling approach is likely to be of value to analyze RCT results. The standard deviation of the agreement scores were also relatively high (SD>1) for these criteria, reflecting the conflicting positions held by panelists. These excluded criteria are described below.

When the outcome rate is lower

A low outcome rate is associated with a more asymmetric distribution of estimated absolute benefit (on the probability scale) across individuals in a trial (Figure 5). For example, when the outcome rate is 6% overall and the C-statistic is 0.80, the average outcome rate in the lowest risk quartile of patients is anticipated to be approximately 1%.[45] While benefit is limited by a floor effect—it is impossible to lower outcome risk to less than zero—

treatment-related harms can still be substantial. The high risk group (e.g. the highest risk quartile) in these low outcome trials frequently has outcome rates more than tenfold the rates found in the low risk group, and may account for most of the benefit in the trial (see Figure 4B). These skewed distributions follow from the logistic regression scale (log odds) and Cox regression scale (log hazard[45]). This makes the average risk (and treatment benefit) misleading even for typical patients enrolled in the trial.[45;90] Nevertheless, there was disagreement among the expert panelists that a low outcome rate was a useful criterion to identify worthwhile target trials for risk modeling. Outcome rate is estimated from empirical data with unavoidable uncertainty and unknown generalizability in other populations, which may have higher outcome rates.

When the two treatments are clinically very different (e.g., medicine versus surgery)

Several treatment selection models have been successfully developed on RCT data comparing treatments that have substantially different mechanisms of action. For example, well known prediction models have been developed on randomized data to disaggregate treatment-favorable from treatment-unfavorable patients for carotid endarterectomy versus medical therapy for patients with symptomatic carotid stenosis[48] or percutaneous coronary intervention (PCI) compared to coronary artery bypass grafting (CABG) for patients with non-acute coronary artery disease.[91] Nevertheless, the TEP disagreed on whether this criterion was a reliable marker of worthwhile opportunities for risk modeling. Indeed, when interventions in alternative trial arms differ substantially, one might anticipate that individual variables may interact with treatment, making an effect modeling approach more advantageous than a risk modeling approach. Indeed, the SYNTAX Score II Model for CABG versus PCI.[91]

Justification of Guidance on Risk Modeling Strategies to Identify HTE (Box B)—The below criteria describe the best methodological practices in the conduct of risk modeling approaches to identify HTE. The motivation regarding these statements are elaborated, including reservations and considerations and caveats.

8. Reporting RCT results stratified by a risk model is encouraged when overall trial results are positive to better understand the distribution of effects across the trial population.

When outcome risk is described using a multivariable model, the control event rate will vary substantially across risk strata of a RCT. It is not uncommon for the control event rate to vary between 5- and even 20-fold across risk strata in trial populations.[45;92] The control event rate is a mathematical determinant of the treatment effect, regardless of what scale is used to measure treatment effect (Table 1). Because typical treatment effect metrics are different (non-linearly related) contrasts of the same two quantities (control and treatment event rates), when the control event rate changes across subgroups, the treatment effect can remain constant on (at most) one scale. In particular, large changes in the control event rate almost always lead to substantial changes in the most clinically relevant scale of effect measure, the absolute risk difference.[4;93] Thus, the widespread assumption that harm-benefit trade-offs are usually similar for patients meeting trial enrollment criteria is demonstrably false and potentially harmfully misleading.[4] Indeed, the assumption of a constant relative treatment effect across groups of patients that vary dramatically in their control event rate —especially in the presence of treatment related harm— needs to be carefully examined. Presenting overall trial results without showing how the treatment effect varies across risk strata—and particularly whether changes in the risk difference are clinically important across risk strata—may be considered a form of under-reporting of trial results.[6]

9. Predictive approaches to HTE require close integration of clinical and statistical reasoning and expertise.

The optimum treatment selection model will generally be grossly underdetermined by the available data, particularly from a single RCT and when multiple potentially important risk markers are considered. Thus, it is generally not possible to use totally agnostic, data-driven approaches alone for variable and model selection when analyzing clinical trials for HTE. Prediction of treatment effect at the individual patient level may be very sensitive to arbitrarily-determined model-building choices that define the reference class (i.e. subgrouping) scheme.[94;95] While in theory, these issues asymptotically diminish as databases become infinitely large, clinical reasoning remains critical to the process of variable selection and model specification in the identification of a clinically-plausible, clinically useful and clinically useable model from the limited data sources generally available. Similarly, given the specialized expertise needed for prediction modeling, clinical investigators should generally not proceed without experienced statistical collaborators. Thus, realizing the goal of predictive HTE analysis requires close partnership between clinical and methodological experts.

Identify or Develop a Model

10. When available, apply a high-quality, externally-developed, compatible risk model to stratify trial results.

For major clinical trials (those that assess a treatment's effect on mortality, major morbidity or other key clinical outcomes), it is often possible to perform risk-based analysis of HTE using an externally developed tool. Prediction models are available to predict overall risk for most major conditions and their complications.[7] Nevertheless, differences in populations or variable definitions may render published models incompatible with completed RCTs. Investigators may also choose to develop a new model using data from a related observational study or clinical trial. An external model is more relevant if the eligibility criteria for the derivation cohort align with, or are even broader than, those in the target trial. Ideally, predictor and outcome variable definitions should be similar to those available in the RCT. An externally-derived model enables translation into practice, especially when well-validated and clinically-accepted models compatible with the RCT are available.

11. 4. When a high-quality, externally-developed model is unavailable, consider developing a model using the entire trial population to stratify trial results; avoid modeling on the control arm only.

When such models are not available, internally derived (or endogenous) models may be employed. Guidance on good prediction modeling practice should be followed, e.g., a large number of events per independent variable, and pre-specified, *a priori,* selection of risk variables based on prior literature.[96;97] Models derived directly on RCT data may provide internally valid treatment effect estimates within risk strata. One approach is to develop the model ignoring treatment assignment.[98] The risk model defines the reference class or subgrouping scheme;[4] a second step then estimates treatment effects across risk strata. Separating the variable selection and model specification process from treatment effect estimation minimizes some of the biases (such as so-called "testimation bias"[96]) that complicate effect modeling. Alternatively, some have recommended that only the control arm be used to model risk[99] as ostensibly the best estimate of the control event rate. However, modeling on the control arm only can potentially induce differential model fit on the two trial arms, biasing treatment effect estimates across risk strata, and generally exaggerating HTE; [46;98;100] various cross-validation techniques have been proposed to address this bias even when modeling on only the control arm.[101] Concerns about differential fit between arms from endogenously derived models may also apply when randomization is unbalanced (e.g., 1:2 or 1:3 randomization) or when treatment effects are very large, such that the number of events in the control arm is much larger than in the treatment arm. When imbalance in events is caused by very large treatment effects, risk-based HTE may be less clinically relevant (see Figure 3). Potential approaches to modeling on trials with imbalanced randomization and/or strong treatment effects should be evaluated in future research (Table 2).

12. 5. When developing new risk models or updating externally-developed risk models, follow guidance for best practice for prediction model development.

While not the focus of the PATH Statement, there is existing guidance to support development of new risk models, or the updating of externally-developed models.[96;97;102;102] In particular, the TRIPOD Explanation and Elaboration document[102] offers detailed, referenced guidance on how to design, conduct, and analyze prediction model studies – with published exemplars – with a view to limiting risk of bias and maximizing the clinical usefulness of the model. We emphasize that TRIPOD indicates that continuous predictors should ideally be kept as continuous (and examined for linear or nonlinear relations with the outcome). This best practice of modeling continuous variables continuously, or modeling risk continuously, does not necessarily proscribe the common practice of displaying clinical trial data to readers in subgroups which is part of the guidance below. The TRIPOD Explanation and Elaboration document[102] also provides a good discussion of the considerations for handling missingness for clinical prediction models, which are relevant here. When important risk factors are missing on some patients, analyses should apply techniques, including multiple imputation when appropriate, to avoid exclusion of randomized patients. Alternative approaches for subgroup identification in the presence of missing variables should be investigated in future research (see Table 4**: A Meta-Research Agenda for Predictive Approaches to Treatment Effect Heterogeneity**). The PROGRESS series of papers[103–106] and several textbooks also offer guidance on the optimal development of clinical prediction models.[96;97]

Since adequate power for HTE analysis might best be achieved by combining multiple randomized trials, prediction modeling guidance should be complemented in these cases by guidance for best practice for individual patient meta-analysis (IPDMA).[107] In particular, it is recommended to include study-specific intercepts to account for unexplained risk heterogeneity.[108;109] Similarly, study-specific effects should be accounted for in analyzing treatment effects. Alternative meta-analytic approaches are discussed in the literature and are beyond the scope of this guidance.[110;111]

Apply the Model and Report Results

13. Report metrics for model performance for outcome risk prediction on the RCT, including measures of discrimination and calibration (when appropriate).

14. Report distribution of predicted risk (or the risk score) in each arm of the trial, and in the overall study population.

15. Report outcome rates and both relative and absolute risk reduction across risk strata.

16. When there are important treatment-related harms, these harms should be reported in each risk strata to support strata-specific evaluation of benefit-harm trade-offs.

Consistent with the TRIPOD Statement, measures of discrimination and calibration should be presented whether an externally or internally derived model is applied (Box B**, Recommendation 6**). However, one should not confuse these conventional measures of model performance with discrimination and calibration of predicted benefit (see Special Considerations for Evaluating Models that Predict Benefit). We also note that point scores, such as TIMI,[112] CHADS-VASC[82] or ABCD2,[113] may be useful for trial risk stratification but do not yield predictions for calibration.

Although its importance was highlighted two decades ago,[92] reporting the distribution of baseline risk is rarely done (Box B**, Recommendation 7**). It is thus generally impossible to assess the degree of baseline risk heterogeneity in most published clinical trials. Risk reporting should allow readers to assess the full distribution of risk in the study population, either graphically or by including information on the mean, standard deviation, median, and interquantile ranges. The precise approach for presentation is not important, as long as it allows the reader to understand the distribution of predicted baseline risk (or the risk score of a risk index) in the study population. The "Table 1" of a clinical trial report (which conventionally includes patient attributes for participants in each study arm) should also include the population mean and median predicted baseline risk (or risk score) with measures of variability, and additional information on the population distribution of risk if there is substantial skewness (such as quartiles/percentiles, a histogram, or a box plot). If the study includes a largely homogeneous population with regard to overall risk, the reader will know that generalizing the study results to populations with substantially different risk would be speculative. If there is substantial heterogeneity in the study population, then reviewers will know that conducting a risk-stratified analysis is particularly important.

We recommend grouping patients using quantiles (e.g. quartiles) for reporting purposes and displaying and estimating treatment effects separately (e.g. dividing patients into equal-sized quarters) in these groups, as an initial step (Box B**, Recommendation 8**). Reporting treatment effect across strata is important because it illustrates how the absolute risk difference varies across the study population, whether or not the *relative* effect is constant (see Figure 4B for an example). Additionally, it permits the assumption of a constant relative effect across risk strata to be evaluated (see Box B**, Recommendation 10,** explanation below). As an alternate presentation, treatment effects may also be displayed by continuous risk, as seen in Figure 2, rather than by quantiles (which are sample-dependent). As discussed above, examining variation in the relative treatment effects may be particularly important when there is even a small amount of treatment-related harm.[3;62] In the case of time-to-event analysis, treatment effects should be analyzed and reported by cumulative incidence curves. Relative treatment effect estimates can be summarized by hazard ratios over a clinically meaningful time horizon (or several meaningful time horizons). Absolute treatment effect estimates can be summarized by cumulative incidences at a clinically meaningful time point (or several meaningful time points). In reporting risk-stratified results, readers should be provided with the information needed to easily determine the amount of variation in risk difference/number needed to treat (NNT) and relative effects. These stratum-specific results can provide a rough guide for clinical interpretation, which can be further refined for clinical implementation by continuous modeling (see Box C**, Recommendation 3**).

From a decision analytic perspective, the clinical value of a prediction model is determined by its ability to distribute patients by their absolute treatment effect across an important decision threshold. This threshold depends on the burdens of treatment, which depend on treatment harms and costs and patient values and preferences. However, even apart from patient values and preferences (which are inherently patient-specific), treatment burden may differ substantially across patient subgroups. Because patients in different risk strata vary in many clinically important characteristics, one cannot assume that subgroups stratified by their risk of the primary outcome have similar rates of treatment-related harms. For example, patients with atrial fibrillation who have higher CHADS scores (indicating higher stroke risk and greater potential benefit from anticoagulation) also have substantially higher risks of bleeding.[69] Stroke patients with a higher risk of stroke recurrence, according to a recurrence risk score, have potentially greater benefit from pioglitazone, but also have a higher risk of pioglitazone-related bone fracture.[61] Because of the potential correlation between these two different risk dimensions (i.e., between the risk of the primary outcome and the risk of treatment-related harm), event rates for these harms should be presented at a congruent level of disaggregation as the primary outcome in order for readers to determine within risk-strata benefit-harm trade-offs (Box B**, Recommendation 9**).

Risk modeling is potentially most useful when predictors of the risk of the primary outcome and the benefits of therapy are poorly (or negatively) correlated with the risks of treatment-related harm. This will maximize heterogeneity in the benefit-harm trade-offs across risk strata, increasing the decisional value of the risk model. While several investigators have sought to arithmetically combine separate prediction models predicting outcome risk and treatment-related harm to stratify trial results by the benefit-harm trade-off,[48;65;68] this approach can be exquisitely sensitive to miscalibration of the two models (which may compound miscalibration of the harm-benefit trade-offs). The best approach to model benefits and harms simultaneously is beyond the scope of these recommendations and is an important topic for future research.

17. To test the consistency of the relative treatment effect across prognostic risk, a continuous measure of risk (e.g., the logit of risk) may be used in an interaction term with treatment group indicator

While it has generally been stressed that testing for a statistical interaction between subgroups is recommended to determine HTE, when the outcome rates vary substantially across strata, one may assume that the risk difference also varies (Table 1; Figure 4). Thus, even though the absolute scale is the most relevant clinically, null hypothesis testing for HTE across risk strata on the risk difference scale is generally not useful, as a non-signficant result is far more likely to reflect low power rather than true consistency of effects on the risk difference scale. Statistically testing a risk-by-treatment interaction on the relative scale (e.g., whether the linear predictor of risk interacts with treatment) provides information on whether a constant relative treatment effect may be a reasonable approximation with which to estimate a risk-specific ("individualized") treatment effect. Yet the presence or absence of a statistically significant result should not be conflated with the clinical significance of HTE (which should always be evaluated on the risk difference scale). In conducting a risk-by-treatment interaction test, using a continuous measure of risk (e.g., the logit of risk) typically provides superior power compared to testing for effect differences across distinct risk groups (see middle panel Figure 4B).[114] Nevertheless, a visual (non-parametric) exploration of how the relative effect varies across values of outcome risk may ensure the appropriateness of linear effect modification. Testing for a non-linear interaction between risk and treatment (e.g., using the logit of risk in a quadratic term, or with another flexible non-linear shape[115;116]) may also be useful. However, it should be recognized that such an interaction test may be poorly powered to detect deviations from linearity, particularly when only a single trial with a limited number of events is the substrate for modeling. Moreover, once it is established that there is an overall treatment effect, determining the risk-specific treatment effect should be considered an estimation problem (rather than a hypothesis testing problem).

Standard errors across levels of risk can be estimated through a proportional interactions model.[117;118] Flexibly modeling the treatment effect across risk strata, or simply reporting the effects across subgroups defined by quantiles (e.g., quartiles) provides useful information regardless of the p-value of the interaction terms testing effect modification on the relative scale. Most importantly, the presence or absence of a statistically significant treatment interaction term should not be conflated with the presence or absence of clinically important HTE (see below).

Justification of Caveats and Considerations Before Moving to Clinical Practice (Box C)—The below considerations relate to the translation of findings from predictive approaches to HTE analyses into clinical practice. Clinical translation of these analyses is a complex topic including many issues, and a detailed discussion of these challenges is beyond the scope of this project, where we focus on analysis and reporting of RCT findings. However, below are some priority analytical considerations for facilitating model translation into clinical practice.

1. Clinical interpretation of HTE should stress differences in the absolute treatment effects across risk groups: the statistical significance of effect modification on the relative scale should not be conflated with the clinical significance of absolute treatment effect estimates.

   As discussed, it is generally agreed that the most important scale for clinical interpretation is the absolute risk difference scale (or its reciprocal, the number needed to treat). The clinical significance of HTE should generally be discussed with reference to the absolute scale, while a relative scale (e.g., odds ratio or hazard ratio) is typically appropriate for null hypothesis testing in HTE analyses. Investigators should consider reporting results in ways that facilitate clinical interpretation of how treatment decisions might be changed with use of the risk model (e.g., number of patients treated or the number of events avoided with versus without model use) and should consider decision analytic approaches for evaluation.[55;119] Table 2 illustrates how results can be presented in a simple way that facilitates understanding of the clinical relevance of risk-modeling. Presentation of important treatment-related harms should also permit within-strata evaluation of absolute effects.

2. External validation and calibration of risk prediction is important for translation of risk-specific treatment effects into clinical practice.

   Although internally-derived (or endogenous) prognostic models can provide reliable *internally valid* estimates of treatment effects within trial risk strata, implementation of an *externally valid* prognostic model is necessary for translation into practice.[105] Finding clinically important HTE across risk strata within a trial with an endogenous model provides an important impetus for developing and implementing an externally valid prognostic model. It should be noted that external validity is a general concern for RCT results and their subgroup analyses, not one confined to results subgrouped using prediction models.[120]

3. Clinical implementation may be supported by translating multivariable risk-based subgroup analysis into models yielding continuous treatment effect predictions to avoid artefactual discontinuities in estimation at the quantile boundary of an outcome risk group.

In presenting HTE analyses of clinical trial results, it is customary to categorize patients into subgroups. Here, we have recommended presenting results in risk strata. Nevertheless, dividing patients into discrete groups based on values of a continuous measure has some disadvantages.[121] Categorization into risk groups suggests that risk and treatment effects are homogeneous within groups, and leads to a potentially misleading "step function" in either risk or in treatment effect estimation. With such an approach, for example, a very small change in risk at the boundary of a group defined by a quantile can lead to a very large change in the anticipated benefit. Additionally, quantiles have specific disadvantages in that they are sample-driven cut points, leading to difficulties in comparing results across studies,[122] and also potentially obscuring problems with model calibration. For example, using an internally developed risk model and the Framingham model to stratify patients in the Diabetes Prevention Program (DPP) Trial appeared to yield near-identical results.[71;123] However, if the trial population was divided into groups of patients based on predicted risk thresholds (as they would in clinical practice), risk groups defined by the Framingham model would have revealed that the Framingham model was poorly calibrated to the DPP Trial population. Thus, while trial results displayed by risk strata are frequently sufficient to evaluate the clinical importance of risk-based HTE, clinical implementation may be supported by translating multivariable risk-based subgroup analysis into models yielding continuous predictions of outcome risks with and without therapy (see Legend Figure 1).

Treatment Effect Modeling to Identify HTE (Box D)—Conventional one-variable-at-a-time subgroup analysis are known to have low credibility due to noisy data (very low power for interactions) weak theory (little prior knowledge about effect modification) and multiplicity[4]. Including relative effect modifiers as interaction terms within a prediction model engenders the same concerns.

While relative effect modifiers are difficult to reliably identify, they are highly influential on individual patient predictions of benefit[120]. Including spurious false positive interaction terms in models that predict treatment benefit can mis-target therapies; excluding true interaction terms limits the usefulness of prediction by substantially lowering discrimination of patients who benefit from those who don't[29]. Whether or not to include a treatment effect interaction term in a prediction model is a fraught and consequential decision; the PATH TEP accordingly recommends a cautious approach. We restrict our recommendations to the unusual situation where highly credible effect modifiers have been identified, and otherwise offer caveats and considerations for more data driven approaches.

1. When highly credible relative effect modifiers have been identified, they should be incorporated into prediction models using multiplicative treatment-by-covariate interaction terms.

A. Credibility should be evaluated using rigorous multidimensional criteria (such as described by the ICEMAN tool) and should not rely solely on statistical criteria (such as p-value thresholds).

There have been important efforts to establish criteria that might identify highly credible subgroup analyses.[9;19;124] A newly proposed tool (Instrument for assessing the Credibility of Effect Modification ANalyses, (ICEMAN)[124]) is based on 5 criteria to evaluate credibility of effect modification: Four of these criteria are related to pre-specification and markers of the "prior probability" or plausibility of effect modification: 1) presence of prior evidence; 2) prespecification of a few primary subgroup analyses; 3) prespecification of the anticipated directionality of effect modification; 4) full specification of cut points when thresholds are used for continuous variables. The fifth criterion (a low p-value) is a measure of the statistical strength of the interaction effect in the data being analyzed. The PATH group endorses the rigorous and multidimensional approach recommended in ICEMAN to identify highly credible interaction terms. Examples of highly credible effect modifiers include symptom onset to treatment time for thrombolytic therapy in acute myocardial infarction or acute ischemic stroke[125;126], and urinary protein excretion as a modifier of the effect of ACE inhibition on the progression of chronic kidney disease[75;127].

In the analysis of trial data, identification of credible interaction terms can be facilitated by explicitly distinguishing those subgroup analyses that are intended to be confirmatory analyses (hypothesis testing analyses, well-motivated by prior evidence and intended to produce clinically actionable results), from secondary (exploratory) subgroup analyses (performed to inform future research).[7;8] Because in any given clinical trial, there is typically limited prior information regarding effect modification, subgroup analyses will frequently be exclusively exploratory, and therefore not yield any covariate-by-treatment interaction effects appropriate for inclusion in prediction models intended for informing clinical care.

Prespecification of primary subgroups should include explicit definitions and categories of the subgroup variables, including cut-off thresholds for continuous or ordinal variables where these are used, and the anticipated direction of the effect modification. By conducting primary subgroup analysis that are few in number, fully pre-specified, hypothesis-driven and statistically robust (i.e. based on multiplicative interactions), subgroups can produce evidence regarding factors that influence the benefit of treatment that might then be carried forward into models to yield clinically actionable predictions.

Although only primary (confirmatory) subgroup analyses are relevant for clinical decisions and the predictive HTE analyses we address here, we acknowledge the importance of secondary subgroup analyses to explore more uncertain or unexpected relationships between individual patient attributes and treatment effects; such analyses are appropriate for hypotheses generation, which can then be tested (and usually disproved[23;25;128;129] in future studies.

We note that because p-values (or other statistical criteria) in general are influential in how subgroup analyses are interpreted (and are included in ICEMAN criteria) and because interaction effects are poorly estimated in trials of conventional size (even when these are pooled), treatment interaction terms selected for inclusion are likely to overestimate the true interaction effects (i.e. from overfitting).[4;29;130] Therefore, even when only highly credible interaction terms are included, we recommend model building procedures that take into account model complexity (i.e. approaches using regularization/penalization) whenever interactions are included (see Box D, Recommendation #3).

Caveats and Considerations for Data Driven Effect Modeling

Emerging ‘data driven’ methods proposed to develop effect models on trial data when there is little prior information on effect modifiers are a promising area of research[28], but were felt by the TEP to be at too formative a stage to offer recommendations. A systematic scoping review[131] was conducted in an effort to characterize this rapidly evolving literature, and revealed future research opportunities. Table 3 summarizes at a high level the key features of the differing approaches. In addition to risk and effect modeling, the scoping review identified methods collectively described as optimal treatment rules that use combinations of relative effect modifiers to classify patients into treatment favorable and treatment unfavorable categories—i.e. based only on the sign of the effect. Because these are classification, rather than prediction methods, we do not discuss these in our guidance, but direct interested readers to the literature on these approaches[132–142]. A key feature of many of the procedures for predictive analyses (and particularly effect modeling) is that they separate the variable selection procedure in building a model that defines subgroups (or reference class scheme) from the estimation of treatment effects, thereby avoiding testimation bias.[96;143;144] Alternatively, various methods of penalization/regularization to reduce the likelihood of overfitting have been proposed and tested.[145] Although the PATH TEP did not articulate a full set of methodological best practices for treatment effect modeling given more limited practical experience (as compared to risk modeling), we offer caveats and considerations for this type of predictive HTE analysis (and their justifications) below.

2. Avoid one-variable-at-a-time null hypothesis testing or stepwise selection (e.g., backward selection, forward selection) strategies to select single relative effect modifiers

One-variable-at-a-time null hypothesis testing will preferentially select effects that are over-estimated within the sample database (i.e., Type I error and testimation bias). Including treatment interaction terms in models predicting benefit generally requires reliable prior information regarding relative effect modifiers. Interaction terms for well-established treatment effect modifiers should be included in the prediction model, regardless of the statistical significance of the interaction (see Box D, Recommendation 1). When multiple relative effect modifiers are hypothesized to be of potential importance in determining treatment effects, the value of including these interactions may be assessed simultaneously by a single overall test, limiting the opportunity for Type I error and testimation bias.[102] To increase the power of this test, the number of treatment effect interactions included should be limited.[102] Investigators should use clinical considerations and should also consider examining associations between candidate effect modifiers to reduce the number of interactions assessed within the overall test. A null result on the overall test for interaction suggests that an effect modeling approach (i.e., including interaction terms) will not add substantially to a risk modeling approach.

3. 3. Avoid the use of regression methods that do not take into account model complexity when estimating coefficients (e.g. ‘conventional’ unpenalized maximum-likelihood regression) when one or more treatment by covariate interaction terms are included in a treatment effect model.

Although there is no consensus on the optimal approach for including relative effect modifiers in a risk prediction model (“effect model”), conventional regression techniques should be anticipated to result in overfitting. Penalized estimation (e.g. LASSO; penalized/ridge; elastic net regularization regression; Bayesian penalization, as well as other non-regression based methods discussed above) at least partially addresses the tendency to overfit benefit predictions. The optimal approach to penalization for effect modeling is a subject of current research, but avoiding overfitting for benefit prediction is much more difficult than avoiding overfitting for outcome prediction (Figure 6). Alternatively, 2-stage methods relying on different datasets (or subsets) for variable and model selection for determining the subgrouping (i.e. reference class) scheme and for treatment effect estimation can also avoid/mitigate ‘testimation’ bias.

4. 4. Avoid evaluating models that predict treatment benefit using only conventional metrics for outcome prediction (e.g. based on discrimination and calibration of outcome risk prediction).

The performance of models intended to predict benefit should be evaluated for the prediction of benefit (i.e., predicted versus observed outcome rates across subgroups defined by quantiles of predicted benefit occurring with *versus* without treatment), not for their ability to predict outcome risk. Calibration for outcome risk can be seriously misleading when evaluating models that purport to predict treatment benefit (Figure 6). The discrepancy arises because benefit miscalibration compounds the error in the risk estimation in the control and treatment groups and magnifies this error (i.e., the scale of the risk difference is typically much smaller than that of outcome rate). Evaluation methods that pertain to models intended for treatment selection or to predict benefit are discussed further in the section immediately below.

Special Considerations for Evaluating Models that Predict Benefit—The statistical performance of prediction models is typically decomposed into measures of calibration (“Do *x* of 100 patients with a predicted risk of x% actually have the outcome?”) and discrimination (“What is the probability that patients with the outcome have a higher predicted risk than those without the outcome?”). Evaluating a prediction model intended to predict treatment effect using these usual metrics related to outcome risk prediction (e.g., C-statistic) fails to provide information on how well the model performs for predicting benefit and informing treatment decisions. Efforts to develop measures to assess model accuracy for predicting benefit (in particular, evaluating measures of discrimination for benefit) are hampered by the fundamental problem of causal inference for the individual. That is, individual patient treatment effects are inherently unobservable, as only one of the possible outcomes is observed for each patient (the actual outcome they experienced under randomization).[146]

For example, for “predictive” biomarkers (i.e., factors that can aid in treatment selection), it has been suggested to evaluate performance by the sensitivity and the specificity of the biomarker for benefit rather than risk, i.e., the probability that the biomarker is positive for patients that benefit from treatment, and the probability that the biomarker is negative for patients that do not benefit from treatment, respectively.[147] However, it has been shown that one can estimate these quantities only under strong unverifiable assumptions about the joint distribution of observed and unobserved outcomes so that each patient can be assigned to a treatment response (bad outcome without treatment, good outcome with treatment, [1,0]), neutral (good or bad outcome regardless of treatment [0,0 or 1,1]) and harm (good outcome without treatment, bad outcome with treatment [0,1])).[148] For example, one proposed method assumes that no subjects are harmed by treatment;[149] others that, conditional on a set of covariates, the potential outcomes with and without treatment are independent.[150;151] Without such assumptions, we can only focus on a model’s ability to predict outcome risk in one arm of a trial or the other, rather than the difference in outcome across arms.[148;152]

However, for treatment decision making purposes, we are more interested in optimizing the ability to predict the difference in outcome risk under two treatment conditions. Due to the fact that counterfactual outcomes are unobservable at the individual patient level, evaluating benefit prediction requires some form of stratification of patients into groups with similar predicted benefit. The smallest possible strata are pairs of matched patients. Recently, the C-statistic, commonly used to measure discrimination in outcome risk models, has been adapted to evaluate treatment effect prediction.[151] To do so, two patients discordant on treatment assignment are matched according to their predicted benefit (i.e., the absolute difference in their outcome risk with and without therapy). These matched pairs of patients with a similar "propensity for benefit" can then be classified into the three benefit categories according to their "observed benefit" based on a comparison of outcomes in the control and treated patients: 1) benefit (1,0); 2) neutral (1,1 or 0,0); or 3) harm (0,1); the C-statistic assesses how well the model discriminates pairs of patients based on this trinary "outcome." Again, the definition of "observed benefit" assumes that the potential outcomes with the two therapies are independent within each patient. Since the potential outcomes within each patient are presumably dependent to some (unknowable) degree, the "observed benefit" contains more randomness than the actual (unobservable) individual patient treatment benefit. This leads to conservative estimates of the C-statistic.

Ultimately, the usefulness of a model depends not just on its ability to provide accurate predictions of within-strata treatment effect, but on its ability to improve decisions. Of course, the ultimate test of a predictive approach is to compare decisions (or outcomes) in settings using such individualized predictions to usual care in an experiment.[153] Lamentably, this is seldom done; well-controlled trials of predictive tools are rare and more are needed. However even in the absence of a randomized trial, methods have been developed to assess the potential impact of models on clinical decision-making. Evaluating clinical usefulness depends on model performance relative to a specific decision threshold—i.e., the absolute risk difference that perfectly balances the burdens, harms, and costs of therapy. Decision curve analysis,[56] has been proposed to evaluate the clinical usefulness of prediction models on decision making. Decision curve analysis examines the Net Benefit across multiple decision thresholds, where each threshold is used to simultaneously determine allocation to a particular treatment strategy and also to mathematically derive an utility weight of benefits versus harms. The approach has also been adapted to evaluate the potential impact of prediction of treatment benefit on decision making compared with the default best overall strategy (i.e., treat all or treat none).[119]

While it is impossible to identify the correct treatment for any given individual (since individual treatment effects are unidentifiable in the absence of repeated N of 1 trials[27;154]), all these methods evaluate whether a particular prediction-decision strategy optimizes benefits for a population,[55] since population-wide benefits are optimized when treatments are optimized for each individual.

Limitations of the PATH Statement—We note several limitations of the PATH Statement. The guidance here is intended only for binary or time to event models, which account for the vast majority of large phase 3 clinical trials.[155–158] Much of the guidance would nevertheless pertain also to continuous outcomes. The complexities of HTE analyses for increasingly common longitudinal data involving interaction with time are not discussed. Further, we focus on treatments where a decision is made at a particular point in time (corresponding to the trial baseline) rather than dynamic treatment regimens where treatment decisions may be continually revisited. We also focus on subgroup identification and treatment effect estimation rather than on HTE analyses to inform trial design. The Statement also does not provide advice about performing n-of-1—or multi-person n-of-1—trials, which some consider the only means of estimating "person-level" treatment effects. While we anticipate that observational studies will play an increasingly important role in studying both treatment effects and HTE, the PATH Statement does not address HTE in observational studies (except to stress that methods of debiasing treatment comparisons to support HTE are a research priority). Although each of the above approaches are consistent with the broad goal of evidence personalization, the methods are sufficiently distinct so as to be beyond the scope of this Statement. Notwithstanding the above limitations, we emphasize that the PATH Statement applies to the comparison of treatments as well as the comparison of treatment to no treatment.

An additional limitation of the PATH Statement is that it does not address the topic of the best estimand for predictive HTE analyses and how to best cope with post-randomization events, including drop out, non-adherence, treatment switching and loss to follow up. These issues have received considerable attention in the methodological and regulatory literature, particularly since The National Research Council (NRC) Expert Panel Report on Prevention and Treatment of Missing Data in Clinical Trials[159;160] highlighted the need for clearly defining objectives and estimands and the subsequent ICH E9(R1) draft addendum[161]. We direct readers to recent literature on this topic[162–164]. In general, the primary analysis of most trials is often an intent to treat (ITT) analysis. However, other contrasts are also of clinical import and interest. In particular, the direct causal effect of treatment (i.e. the effect *if* a patient adheres to their treatment, estimated with a per protocol (PP) or adherence-adjusted analysis[165–169]) is often considered the most appropriate estimand for shared decision making in the individual patient. However, as with observational studies, estimating the direct treatment effect can only be done with methods based on unverifiable assumptions; mis-specifying a model predicting non-adherence (or using an instrumental variable approach when causes of non-adherence or drop outs are complex) can lead to biased estimates of treatment effects. An ITT analysis is general felt to yield an unbiased estimate of the treatment policy, though this may be less appropriate for shared decision making. More research is needed regarding optimal approaches to combine predictive HTE approaches with approaches that estimate direct treatment or adherence-adjusted effects.

Discussion

The PATH Statement is comprised of four sets of guidance on the conduct of predictive HTE analyses. The purpose of the Explanation and Elaboration document is to explain the rationale and support for these guidance statements, and to detail caveats or reservations where applicable.

The goal of predictive HTE analysis is prediction of treatment effect to support decision making in each patient.[8;170] Developers of the PATH Statement recognize the inherent difficulties and fundamental limitations of using group data to estimate treatment effects in individuals, and enumerated some of these challenges.[26] As more deeply explored in this Explanation and Elaboration document, there remain substantial barriers to fully understand the potential of the predictive HTE approaches to usher in a new era of Personalized EBM.[171] Table 4 outlines some outstanding research questions related to methodological issues raised in the development of the PATH Statement. Stronger methodological and evidentiary standards will need to be established to assure that incorporating these methods do not cause more harm than benefit. We also need research to better integrate clinical prediction into practice,[172] to understand how to individualize clinical practice guidelines, to establish or extend reporting guidelines,[173] to establish new models of data ownership to facilitate individual level meta-analyses,[174] and to re-engineer the clinical research infrastructure to support substantially larger clinically-integrated trials sufficiently powered to determine HTE (and/or to develop our ability to predict when observational data are likely to be sufficiently de-biased for reliable HTE determination).[175] Many recent and ongoing organizational and technical advances should help towards this evolution.[174;176–178]

As the PATH Statement focused on prediction in randomized trials, we did not explore the use of observational data – and when they may be sufficiently de-biased for reliably identifying HTE.[10;179;180] In addition, there is an evolving set of tools for data-driven approaches to predicting patient benefit, including machine learning techniques.[28;181;182] The PATH Statement should thus be understood as a formative first step (along a much longer path) towards the goal of personalized predictions of treatment benefit using the best available evidence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We acknowledge the excellent technical support from Mark Adkins, Teddy Balan, Jeff Chang, and Dan Sjoberg, in analyses included in figures and supporting appendices. We also thank the Annals of Internal Medicine editors and reviewers, whose thoughtful feedback greatly improved this work. We thank Jennifer Lutz and Christine Lundquist for assistance with copyediting and creating exhibits.

Reference List

(1). Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet 1995; 345(8965):1616–1619. [PubMed: 7783541]

(2). Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet 2005; 365(9455):256–265. [PubMed: 15652609]

(3). Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007; 298(10):1209–1212. [PubMed: 17848656]

(4). Kent DM, Steyerberg EW, van Klaveren D. Personalized evidence-based medicine: predictive approaches to heterogeneous treatment effects. BMJ 2018; 363:k4245.

(5). Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004; 82(4):661–687. [PubMed: 15595946]

(6). Hayward RA, Kent DM, Vijan S, Hofer TP. Reporting clinical trial results to inform providers, payers, and consumers. Health Aff (Millwood) 2005; 24(6):1571–1581. [PubMed: 16284031]

(7). Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010; 11:85. [PubMed: 20704705]

(8). Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2013; 66(8):818–825. [PubMed: 23651763]

(9). Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014; 311(4):405–411. [PubMed: 24449319]

(10). Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. Int J Epidemiol 2016; 45(6):2184–2193. [PubMed: 27864403]

(11). Davidoff F. Can Knowledge About Heterogeneity in Treatment Effects Help Us Choose Wisely? Ann Intern Med 2017; 166(2):141–142. [PubMed: 27820948]

(12). Lagakos SW. The challenge of subgroup analyses--reporting without distorting. N Engl J Med 2006; 354(16):1667–1669. [PubMed: 16625007]

(13). Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005; 365(9454):176–186. [PubMed: 15639301]

(14). Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006; 151(2):257–264. [PubMed: 16442886]

(15). Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine--reporting of subgroup analyses in clinical trials. N Engl J Med 2007; 357(21):2189–2194. [PubMed: 18032770]

(16). Furberg CD, Byington RP. What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-Blocker Heart Attack Trial experience. Circulation 1983; 67(6 Pt 2):I98–101. [PubMed: 6133654]

(17). Tannock IF. False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. J Natl Cancer Inst 1996; 88(3–4):206–207. [PubMed: 8632495]

(18). Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355(9209):1064–1069. [PubMed: 10744093]

(19). Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116(1):78–84. [PubMed: 1530753]

(20). Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002; 21(19):2917–2930. [PubMed: 12325108]

(21). Stallones RA. The use and abuse of subgroup analysis in epidemiological research. Prev Med 1987; 16(2):183–194. [PubMed: 3295858]

(22). Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000; 139(6):952–961. [PubMed: 10827374]

(23). Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey SG. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001; 5(33):1–56.

(24). Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol 2004; 57(3):229–236. [PubMed: 15066682]

(25). Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015; 351:h5651.

(26). Kent DM, Paulus JK, D'Agostino R, Goodman S, Hayward R, Ioannidis JPA et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. AIM 2018.

(27). Raman G, Balk EM, Lai L, Shi J, Chan J, Lutz JS et al. Evaluation of person-level heterogeneity of treatment effects in published multi-person N-of-1 studies: systematic review and re-analysis. BMJ Open 2018; 8(5):e017641.

(28). Lipkovich I, Dmitrienko A, D'Agostino BR Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 2017; 36(1):136–196. [PubMed: 27488683]

(29). van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol 2019; 114:72–83. [PubMed: 31195109]

(30). Greenland S, Lash TL, Rothman KJ. Concepts of Interaction In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. 3rd ed Philadelphia: Lippincott Williams and Wilkins; 2008.

(31). VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. Epidemiology 2007; 18(3):329–339. [PubMed: 17435441]

(32). VanderWeele TJ, Knol MJ. A tutorial on interaction. Epidemiologic Methods 2014; 3(1):33–72.

(33). Hallqvist J, Ahlbom A, Diderichsen F, Reuterwall C. How to evaluate interaction between causes: a review of practices in cardiovascular epidemiology. J Intern Med 1996; 239(5):377–382. [PubMed: 8642229]

(34). Andersson T, Alfredsson L, Kallberg H, Zdravkovic S, Ahlbom A. Calculating measures of biological interaction. Eur J Epidemiol 2005; 20(7):575–579. [PubMed: 16119429]

(35). Ahlbom A, Alfredsson L. Interaction: A word with two meanings creates confusion. Eur J Epidemiol 2005; 20(7):563–564. [PubMed: 16119427]

(36). VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology 2007; 18(5):561–568. [PubMed: 17700242]

(37). VanderWeele TJ, Robins JM. Empirical and counterfactual conditions for sufficient cause interactions. Biometrika 2008; 95:4961.

(38). VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. Ann Intern Med 2011; 154(10):680–683. [PubMed: 21576536]

(39). Harrell F, Lazzeroni L. EHRs and RCTs: Outcome prediction vs. optimal treatment selection. 2017 Available from: http://www.fharrell.com/post/ehrs-rcts/.

(40). Harrell F. Viewpoints on heterogeneity of treatment effect and precision medicine. 2018 Available from: http://www.fharrell.com/post/hteview/.

(41). Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med 2000; 19(13):1707–1728. [PubMed: 10861773]

(42). Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001; 54(10):1046–1055. [PubMed: 11576817]

(43). Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes In: Systematic Reviews in Health Care: Meta-Analysis in Context M Egger G Davey Smith; and Altman DG(Eds), 313335. London: BMJ Publishing Group; 2003.

(44). Goff DC Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr., Gibbons R et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2014; 63(25 Pt B):2935–2959. [PubMed: 24239921]

(45). Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol 2016; 1(45):2075–2088.

(46). van Klaveren D, Balan TA, Steyerberg EW, Kent DM. Models with interactions overestimated the heterogeneity of treatment effect in simulated trials. Journal of Clinical Epidemiology 2018.

(47). Kent DM, Ruthazer R, Selker HP. Are some patients likely to benefit from recombinant tissue-type plasminogen activator for acute ischemic stroke even beyond 3 hours from symptom onset? Stroke 2003; 34(2):464–467. [PubMed: 12574561]

(48). Rothwell PM, Warlow CP. Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group. Lancet 1999; 353(9170):2105–2110. [PubMed: 10382694]

(49). Frommeyer G, Eckardt L. Drug-induced proarrhythmia: risk factors and electrophysiological mechanisms. Nat Rev Cardiol 2016; 13(1):36–47. [PubMed: 26194552]

(50). Roden DM. Mechanisms and management of proarrhythmia. Am J Cardiol 1998; 82(4A):49I–57I.

(51). Costa F, van Klaveren D., James S, Heg D, Raber L, Feres F et al. Derivation and validation of the predicting bleeding complications in patients undergoing stent implantation and subsequent dual antiplatelet therapy (PRECISE-DAPT) score: a pooled analysis of individual-patient datasets from clinical trials. Lancet 2017; 389(10073):1025–1034. [PubMed: 28290994]

(52). Viscoli CM, Kent DM, Conwit R, Dearborn JL, Furie KL, Gorman M et al. A scoring system to optimize pioglitazone therapy after stroke based on fracture risk. 2018.

(53). Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc 2009; 172(4):729–748.

(54). Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980; 302(20):1109–1117. [PubMed: 7366635]

(55). Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016; 352:i6.

(56). Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006; 26(6):565–574. [PubMed: 17099194]

(57). Baker SG, Kramer BS. Evaluating a new marker for risk prediction: decision analysis to the rescue. Discov Med 2012; 14(76):181–188. [PubMed: 23021372]

(58). Hammadah M, Kim JH, Tahhan AS, Kindya B, Liu C, Ko YA et al. Use of High-Sensitivity Cardiac Troponin for the Exclusion of Inducible Myocardial Ischemia: A Cohort Study. Ann Intern Med 2018; 169(11):751–760. [PubMed: 30398528]

(59). Baker SG. Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst 2009; 101(22):1538–1542. [PubMed: 19843888]

(60). Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM et al. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 2008; 167(3):362–368. [PubMed: 17982157]

(61). Yao X, Gersh BJ, Sangaralingham LR, Kent DM, Shah ND, Abraham NS et al. Comparison of the CHA2DS2-VASc, CHADS2, HAS-BLED, ORBIT, and ATRIA Risk Scores in Predicting Non-Vitamin K Antagonist Oral Anticoagulants-Associated Bleeding in Patients With Atrial Fibrillation. Am J Cardiol 2017; 120(9):1549–1556. [PubMed: 28844514]

(62). Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol 2006; 6:18. [PubMed: 16613605]

(63). Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA 2001; 285(22):2864–2870. [PubMed: 11401607]

(64). Gage BF, van Walraven C, Pearce L, Hart RG, Koudstaal PJ, Boode BS et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation 2004; 110(16):2287–2292. [PubMed: 15477396]

(65). Kent DM, Hayward RA, Griffith JL, Vijan S, Beshansky JR, Califf RM et al. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. Am J Med 2002; 113(2):104–111. [PubMed: 12133748]

(66). Thune JJ, Hoefsten DE, Lindholm MG, Mortensen LS, Andersen HR, Nielsen TT et al. Simple risk stratification at admission to identify patients with reduced mortality from primary angioplasty. Circulation 2005; 112(13):2017–2021. [PubMed: 16186438]

(67). Fox KA, Poole-Wilson P, Clayton TC, Henderson RA, Shaw TR, Wheatley DJ et al. 5-year outcome of an interventional strategy in non-ST-elevation acute coronary syndrome: the British Heart Foundation RITA 3 randomised trial. Lancet 2005; 366(9489):914–920. [PubMed: 16154018]

(68). Yeh RW, Secemsky EA, Kereiakes DJ, Normand SL, Gershlick AH, Cohen DJ et al. Development and Validation of a Prediction Rule for Benefit and Harm of Dual Antiplatelet Therapy Beyond 1 Year After Percutaneous Coronary Intervention. JAMA 2016; 315(16):1735–1749. [PubMed: 27022822]

(69). Kernan WN, Viscoli CM, Dearborn JL, Kent DM, Conwit R, Fayad P et al. Targeting Pioglitazone Hydrochloride Therapy After Stroke or Transient Ischemic Attack According to Pretreatment Risk for Stroke or Myocardial Infarction. JAMA Neurol 2017; 174(11):1319–1327.

(70). Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002; 346(6):393–403. [PubMed: 11832527]

(71). Sussman JB, Kent DM, Nelson JP, Hayward RA. Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. BMJ 2015; 350:h454.

(72). Kent DM, Vijan S, Hayward RA, Griffith JL, Beshansky JR, Selker HP. Tissue plasminogen activator was cost-effective compared to streptokinase in only selected patients with acute myocardial infarction. J Clin Epidemiol 2004; 57(8):843–852. [PubMed: 15485737]

(73). Ioannidis JP, Garber AM. Individualized cost-effectiveness analysis. PLoS Med 2011; 8(7):e1001058.

(74). Schuit E, Li AH, Ioannidis JPA. How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study. Int J Epidemiol 2018.

(75). Kent DM, Jafar TH, Hayward RA, Tighiouart H, Landa M, de JP et al. Progression risk, urinary protein excretion, and treatment effects of angiotensin-converting enzyme inhibitors in nondiabetic kidney disease. J Am Soc Nephrol 2007; 18(6):1959–1965. [PubMed: 17475813]

(76). Trikalinos TA, Ioannidis JP. Predictive modeling and heterogeneity of baseline risk in meta-analysis of individual patient data. J Clin Epidemiol 2001; 54(3):245–252. [PubMed: 11223322]

(77). van Klaveren D, Gonen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Stat Med 2016; 35(23):4136–4152. [PubMed: 27251001]

(78). Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. Journal of Clinical Epidemiology 1997; 50(10):1089–1098. [PubMed: 9368516]

(79). Wessler BS, Paulus JK, Lundquist C, Ajlan M, Natto Z, Janes WA et al. Tufts PACE Clinical Prediction Model Registry: update 1990 through 2015. Diagnostic and Prognostic Research 2017; 1(20):1–8. [PubMed: 31093533]

(80). Wessler BS, Lai YL, Kramer W, Cangelosi M, Raman G, Lutz JS et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circ Cardiovasc Qual Outcomes 2015; 8(4):368–375. [PubMed: 26152680]

(81). Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA 2018; 320(1):27–28. [PubMed: 29813156]

(82). Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010; 137(2):263–272. [PubMed: 19762550]

(83). Selker HP, Beshansky JR, Griffith JL, Aufderheide TP, Ballin DS, Bernard SA et al. Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial. AIM 1998; 129(11):845–855.

(84). Hess EP, Hollander JE, Schaffer JT, Kline JA, Torres CA, Diercks DB et al. Shared decision making in patients with low risk chest pain: prospective randomized pragmatic trial. BMJ 2016; 355:i6165.

(85). Stiell IG, Clement CM, McKnight RD, Brison R, Schull MJ, Rowe BH et al. The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med 2003; 349(26):2510–2518. [PubMed: 14695411]

(86). Kuppermann N, Holmes JF, Dayan PS, Hoyle JD Jr., Atabaki SM, Holubkov R et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet 2009; 374(9696):1160–1170. [PubMed: 19758692]

(87). Wells PS, Anderson DR, Rodger M, Stiell I, Dreyer JF, Barnes D et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. Ann Intern Med 2001; 135(2):98–107. [PubMed: 11453709]

(88). de Koning L, Merchant AT, Pogue J, Anand SS. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. Eur Heart J 2007; 28(7):850–856. [PubMed: 17403720]

(89). Vazquez G, Duval S, Jacobs DR Jr., Silventoinen K. Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. Epidemiol Rev 2007; 29:115–128. [PubMed: 17494056]

(90). Vickers AJ, Kent DM. The Lake Wobegon effect: Why most patients are at below-average risk. Ann Intern Med 2015; 162(12):886–867.

(91). Farooq V, van Klaveren D, Steyerberg EW, Meliga E, Vergouwe Y, Chieffo A et al. Anatomical and clinical characteristics to guide decision making between coronary artery bypass surgery and percutaneous coronary intervention for individual patients: development and validation of SYNTAX score II. Lancet 2013; 381(9867):639–650. [PubMed: 23439103]

(92). Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. Am J Epidemiol 1998; 148(11):1117–1126. [PubMed: 9850135]

(93). Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2018; 100:22–31. [PubMed: 29654822]

(94). Stern RH. Individual risk. J Clin Hypertens (Greenwich) 2012; 14(4):261–264. [PubMed: 22458749]

(95). Kent DM, Shah ND. Risk models and patient-centered evidence: should physicians expect one right answer? JAMA 2012; 307(15):1585–1586. [PubMed: 22511683]

(96). Steyerberg EW. Clinical prediction models: a practical approach to development, valdiation, and updating. New York: Springer; 2009.

(97). Harrell FE. Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis, second edition NY: Springer; 2015.

(98). Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes 2014; 7(1):163–169. [PubMed: 24425710]

(99). Wang R, Lagakos SW. Response to letter to the editor: “More on subgroup analyses in clinical trials”. N Engl J Med 2008; 358:2076–2077. [PubMed: 18463389]

(100). Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experients (12 2013). Working Paper No.w19742. 2013. Available from: http://ssrn.com/abstract=2370198.

(101). Verver D, van KD, van Akkooi ACJ, Rutkowski P, Powell BWEM, Robert C et al. Risk stratification of sentinel node-positive melanoma patients defines surgical management and adjuvant therapy treatment considerations. Eur J Cancer 2018; 96:25–33. [PubMed: 29660597]

(102). Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015; 162(1):W1–73.

(103). Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A et al. Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ 2013; 346:e5595.

(104). Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA et al. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. PLoS Med 2013; 10(2):e1001380.

(105). Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLoS Med 2013; 10(2):e1001381.

(106). Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW et al. Prognosis research strategy (PROGRESS) 4: Stratified medicine research. BMJ 2013; 346:e5793.

(107). Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA 2015; 313(16):1657–1665. [PubMed: 25919529]

(108). Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med 2013; 32(18):3158–3180. [PubMed: 23307585]

(109). Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol 2014; 14:3. [PubMed: 24397587]

(110). Turner RM, Omar RZ, Yang M, Goldstein H, Thompson SG. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med 2000; 19(24):3417–3432. [PubMed: 11122505]

(111). Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 2010; 340:c221.

(112). Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G et al. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. JAMA 2000; 284(7):835–842. [PubMed: 10938172]

(113). Johnston SC, Rothwell PM, Nguyen-Huynh MN, Giles MF, Elkins JS, Bernstein AL et al. Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. Lancet 2007; 369(9558):283–292. [PubMed: 17258668]

(114). Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995; 6(4):450–454. [PubMed: 7548361]

(115). Harrell FE. Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. NY: Springer; 2001.

(116). Royston P, Sauerbrei W. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chinester: John Wiley & Songs Ltd; 2008.

(117). Kovalchik SA, Varadhan R, Weiss CO. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Stat Med 2013; 32(28):4906–4923. [PubMed: 23788362]

(118). Follmann DA, Proschan MA. A multivariate test of interaction for use in clinical trials. Biometrics 1999; 55(4):1151–1155. [PubMed: 11315061]

(119). Vickers AJ, Kattan MW, Daniel S. Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials 2007; 8:14. [PubMed: 17550609]

(120). van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol 2015; 68(11):1366–1374. [PubMed: 25814403]

(121). Weinberg CR. How bad is categorization? Epidemiology 1995; 6(4):345–347. [PubMed: 7548338]

(122). Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol 2012; 12:21. [PubMed: 22375553]

(123). Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med 2007; 167(10):1068–1074. [PubMed: 17533210]

(124). Instrument for assessing the credibility of effect modification analyses (ICEMAN) in a meta-analysis of randomized controlled trials [Under review at Annals of Internal Medicine].

(125). Hacke W, Donnan G, Fieschi C, Kaste M, von Kummer R, Broderick JP et al. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. Lancet 2004; 363(9411):768–774. [PubMed: 15016487]

(126). Emberson J, Lees KR, Lyden P, Blackwell L, Albers G, Bluhmki E et al. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. Lancet 2014; 384(9958):1929–1935. [PubMed: 25106063]

(127). Jafar TH, Stark PC, Schmid CH, Landa M, Maschio G, Marcantoni C et al. Proteinuria as a modifiable risk factor for the progression of non-diabetic renal disease. Kidney Int 2001; 60(3):1131–1140. [PubMed: 11532109]

(128). Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials. JAMA Intern Med 2017; 177(4):554–560. [PubMed: 28192563]

(129). Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JP. Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. BMJ 2016; 355:i5826.

(130). Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008; 19(5):640–648. [PubMed: 18633328]

(131). Paulus JK, Raman G, Rekkas A, Koethe B, Tanprasertsuk J, Lutz JS et al. White Paper, Appendix 1: Methods and Results of Evidence Review Committee Search The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Washington, D.C: Patient-Centered Outcomes Research Institute (PCORI); 2019.

(132). Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat 2011; 39(2):1180–1210. [PubMed: 21666835]

(133). Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics 2012; 68(4):1010–1018. [PubMed: 22550953]

(134). Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat 2012; 1(1):103–114. [PubMed: 23645940]

(135). Kraemer HC. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Stat Med 2013; 32(11):1964–1973. [PubMed: 23303653]

(136). Wallace ML, Frank E, Kraemer HC. A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. JAMA Psychiatry 2013; 70(11):1241–1247. [PubMed: 24048258]

(137). Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates. J Am Stat Assoc 2014; 109(508):1517–1532. [PubMed: 25729117]

(138). Taylor JMG, Chang W, Foster JC. Reader reaction to “a robust method for estimating optimal treatment regimes” by Zhang et al. (2012). Biometrics 2015; 71(1):267–273.

(139). Xu Y, Yu M, Zhao YQ, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics 2015; 71(3):645–653. [PubMed: 25962845]

(140). Foster JC, Taylor JM, Kaciroti N, Nan B. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. Biostatistics 2015; 16(2):368–382. [PubMed: 25398774]

(141). Niles AN, Loerinc AG, Krull JL, Roy-Byrne P, Sullivan G, Sherbourne CD et al. Advancing Personalized Medicine: Application of a Novel Statistical Method to Identify Treatment Moderators in the Coordinated Anxiety Learning and Management Study. Behav Ther 2017; 48(4):490–500. [PubMed: 28577585]

(142). Petkova E, Tarpey T, Su Z, Ogden RT. Generated effect modifiers (GEM’s) in randomized clinical trials. Biostatistics 2017; 18(1):105–118. [PubMed: 27465235]

(143). Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 2011; 12(2):270–282. [PubMed: 20876663]

(144). Claggett B, Tian L, Castagno D, Wei LJ. Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. Biostatistics 2015; 16(1):60–72. [PubMed: 25122189]

(145). Basu S, Sussman JB, Rigdon J, Steimle L, Denton BT, Hayward RA. Benefit and harm of intensive blood pressure treatment: Derivation and validation of risk models using data from the SPRINT and ACCORD trials. PLoS Med 2017; 14(10):e1002410.

(146). Djulbegovic B, Ioannidis JPA. Precision medicine for individual patients should use population group averages and larger, not smaller, groups. Eur J Clin Invest 2018;e13031.

(147). Sensitivity Simon R., Specificity PPV, and NPV for Predictive Biomarkers. J Natl Cancer Inst 2015; 107(8).

(148). Janes H, Pepe MS, McShane LM, Sargent DJ, Heagerty PJ. The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment. J Natl Cancer Inst 2015; 107(8).

(149). Zhang Z, Nie L, Soon G, Liu A. The Use of Covariates and Random Effects in Evaluating Predictive Biomarkers Under a Potential Outcome Framework. Ann Appl Stat 2014; 8(4):2336–2355. [PubMed: 26779295]

(150). Huang Y, Gilbert PB, Janes H. Assessing treatment-selection markers using a potential outcomes framework. Biometrics 2012; 68(3):687–696. [PubMed: 22299708]

(151). van Klaveren D, Steyerberg EW, Serruys PW, Kent DM The proposed ‘concordance-statistic for benefit’ provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol 2018; 94:59–68. [PubMed: 29132832]

(152). Fine JP, Pencina M. On the quantitative assessment of predictive biomarkers. J Natl Cancer Inst 2015; 107(8).

(153). Selker HP, Beshansky JR, Griffith JL, TPI T, I. Use of the electrocardiograph-based thrombolytic predictive instrument to assist thrombolytic and reperfusion therapy for acute myocardial infarction. A multicenter, randomized, controlled, clinical effectiveness trial. Ann Intern Med 2002; 137(2):87–95. [PubMed: 12118963]

(154). Senn S. Individual response to treatment: is it a valid assumption? BMJ 2004; 329(7472):966–968. [PubMed: 15499115]

(155). Song SY, Seo H, Kim G, Kim AR, Kim EY. Trends in endpoint selection in clinical trials of advanced breast cancer. J Cancer Res Clin Oncol 2016; 142(11):2403–2413. [PubMed: 27586374]

(156). Ghimire S, Kyung E, Kim E. Reporting trends of outcome measures in phase II and phase III trials conducted in advanced-stage non-small-cell lung cancer. Lung 2013; 191(4):313–319. [PubMed: 23715997]

(157). Goldfarb M, Drudi L, Almohammadi M, Langlois Y, Noiseux N, Perrault L et al. Outcome Reporting in Cardiac Surgery Trials: Systematic Review and Critical Appraisal. J Am Heart Assoc 2015; 4(8):e002204.

(158). Phillips R, Hazell L, Sauzet O, Cornelius V. Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open 2019; 9(2):e024537.

(159). National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: The National Academies Press; 2010.

(160). Little RJ, D’Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012; 367(14):1355–1360. [PubMed: 23034025]

(161). International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. Addendum: Statistical Principles for Clinical Trials - Estimands and Sensitivity Analysis in Clinical Trials: Current Step 2 version.2017.E9(R1)

(162). Hernán MA, Scharfstein D. Cautions as Regulators Move to End Exclusive Reliance on Intention to TreatCautions as Regulators Move to End Exclusive Reliance on Intention to Treat. AIM 2018; 168(7):515–516.

(163). Ratitch B, Bell J, Mallinckrodt C, Bartlett JW, Goel N, Molenberghs G et al. Choosing Estimands in Clinical Trials: Putting the ICH E9(R1) Into Practice. Drug Inf J 2019;2168479019838827.

(164). Mallinckrodt C, Bell J, Liu G, Ratitch B, O’Kelly M, Lipkovich I et al. Aligning Estimators With Estimands in Clinical Trials: Putting the ICH E9(R1) Guidelines Into Practice. Ther Innov Regul Sci 2019.

(165). Sussman JB, Hayward RA. An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials. BMJ 2010; 340:c2073.

(166). Bond SJ, White IR, Sarah Walker A. Instrumental variables and interactions in the causal analysis of a complex clinical trial. Stat Med 2007; 26(7):1473–1496. [PubMed: 16900567]

(167). Sommer A, Zeger SL. On estimating efficacy from clinical trials. Stat Med 1991; 10(1):45–52. [PubMed: 2006355]

(168). Hernán MA, Hernández-Diaz S, Robins JM. Randomized Trials Analyzed as Observational Studies. Ann Intern Med 2013; 159(8):560–562. [PubMed: 24018844]

(169). Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med 2017; 377(14):1391–1398. [PubMed: 28976864]

(170). Byar DP. Assessing apparent treatment--covariate interactions in randomized clinical trials. Stat Med 1985; 4(3):255–263. [PubMed: 4059716]

(171). Salisbury AC, Spertus JA. Realizing the Potential of Clinical Risk Prediction Models: Where Are We Now and What Needs to Change to Better Personalize Delivery of Care? Circ Cardiovasc Qual Outcomes 2015; 8(4):332–334. [PubMed: 26152684]

(172). Decker C, Garavalia L, Garavalia B, Gialde E, Yeh RW, Spertus J et al. Understanding physician-level barriers to the use of individualized risk estimates in percutaneous coronary intervention. Am Heart J 2016; 178:190–197. [PubMed: 27502869]

(173). Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015; 131(2):211–219. [PubMed: 25561516]

(174). Krumholz HM, Ross JS, Gross CP, Emanuel EJ, Hodshon B, Ritchie JD et al. A historic moment for open science: the Yale University Open Data Access project and medtronic. Ann Intern Med 2013; 158(12):910–911. [PubMed: 23778908]

(175). Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? JAMA 2014; 312(2):129–130. [PubMed: 25005647]

(176). Vickers AJ, Scardino PT. The clinically-integrated randomized trial: proposed novel method for conducting large trials at low cost. Trials 2009; 10:14. [PubMed: 19265515]

(177). van Staa TP, Klungel O, Smeeth L. Use of electronic healthcare records in large-scale simple randomized trials at the point of care for the documentation of value-based medicine. J Intern Med 2014; 275(6):562–569. [PubMed: 24635449]

(178). Fiore LD, Lavori PW. Integrating Randomized Comparative Effectiveness Research with Patient Care. N Engl J Med 2016; 374(22):2152–2158. [PubMed: 27248620]

(179). Franklin JM, Schneeweiss S. When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials? Clin Pharmacol Ther 2017; 102(6):924–933. [PubMed: 28836267]

(180). Byar DP. Why data bases should not replace randomized clinical trials. Biometrics 1980; 36(2):337–342. [PubMed: 7407321]

(181). Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015; 349(6245):255–260. [PubMed: 26185243]

(182). LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521(7553):436–444. [PubMed: 26017442]

(183). Iwashyna TJ, Burke JF, Sussman JB, Prescott HC, Hayward RA, Angus DC. Implications of Heterogeneity of Treatment Effect for Reporting and Analysis of Randomized Trials in Critical Care. Am J Respir Crit Care Med 2015; 192(9):1045–1051. [PubMed: 26177009]

(184). Groenwold RH, Moons KG, Pajouheshnia R, Altman DG, Collins GS, Debray TP et al. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol 2016; 78:90–100. [PubMed: 27045189]

(185). Abadie A, Chingos MM, West MR. Endogenous stratification in randomized experiments. The Review of Economics and Statistics 2018; 100(4):567–580.

(186). Weisberg HI, Pontes VP. Post hoc subgroups in clinical trials: Anathema or analytics? Clin Trials 2015; 12(4):357–364. [PubMed: 26062595]

(187). Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future compartive study. J Am Stat Assoc 2013; 108(502):527–539. [PubMed: 24058223]

(188). Julien M, Hanley JA. Profile-specific survival estimates: making reports of clinical trials more patient-relevant. Clin Trials 2008; 5(2):107–115. [PubMed: 18375648]

(189). Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ 2011; 343:d5888.

(190). Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. J Biopharm Stat 2014; 24(1):110–129. [PubMed: 24392981]

(191). Chen W, Ghosh D, Raghunathan TE, Sargent DJ. Bayesian variable selection with joint modeling of categorical and survival outcomes: an application to individualizing chemotherapy treatment in advanced colorectal cancer. Biometrics 2009; 65(4):1030–1040. [PubMed: 19210736]

(192). Gunter L, Zhu J, Murphy S. Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. J Biopharm Stat 2011; 21(6):1063–1078. [PubMed: 22023676]

(193). Ternes N, Rotolo F, Heinze G, Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. Biom J 2017; 59(4):685–701. [PubMed: 27862181]

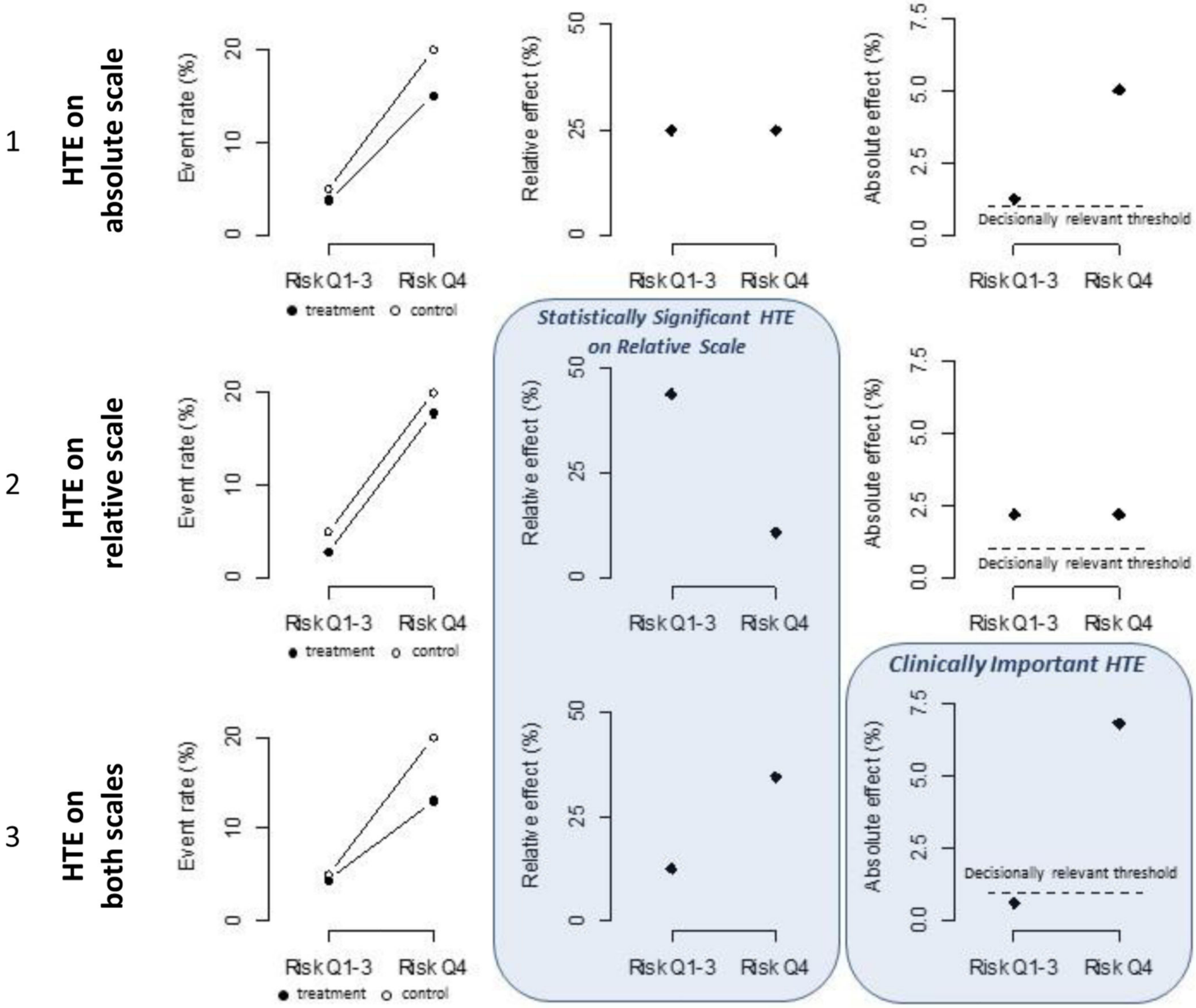


Figure 1. The Scale Dependence of Heterogeneity of Treatment Effect (HTE)

The above plots depict the scale dependence of effect heterogeneity. All 3 scenarios are drawn from hypothetical trials with the same overall results (outcome rates 8.8% versus 6.6% in the control (open circles) versus treatment (closed circles) groups) and depict outcomes in low risk (75% of patients, Q1–3) and high risk (25% of patients, Q4) groups (where control event rates are 5% versus 20%, respectively). Plots in the left, middle and right column display outcome risks, relative effects and absolute effects, respectively. In the first row, effect heterogeneity is absent on the relative scale, but present on the absolute scale. In the second row, effect heterogeneity is present on the relative scale but absent on the absolute scale. In third row, effect heterogeneity is present on both the relative and the absolute scale. Most typically, the statistical significance of HTE is tested on the relative scale (middle column), since regression analyses are often performed on these scales. Provided sufficient statistical power, analyses 2 and 3 would show statistically significant HTE. However, regardless of the scale of the analysis, the clinical importance of HTE should generally be evaluated on the absolute scale. When absolute effects span a decisionally-important threshold, which depends on the treatment burden (e.g. harms and costs), HTE is said to be clinically important. In this example, for illustratrive purposes we have arbitrarily set a decisionally relevant threshold at a 1 percentage point reduction in outcome risk. Here, while there is HTE on the absolute scale in both analyses 1 and 3, clinically-important heterogeneity is present only in the third analysis, where the treatment that is beneficial on average may not be worth the treatment burden for many (indeed most) patients. Note, the presence of statistically significant interaction (on the relative scale) does not imply clinically important HTE, and that the absence of a statistically significant interaction does not imply the absence of clinically important HTE. It is also important to note that testing heterogeneity on the relative scale does not test a specific causal hypothesis regarding effect modification (regardless of the subgrouping variable), but merely tests the hypothesis that relative effects are the same in one group versus another group. Establishing causal interaction effects are not necessary to improve the targeting of therapy. We also note that this diagram makes the simplifying assumption of uniform treatment burdens across all levels of risk.In practice adverse events may vary across risk groups, and the threshold is also sensitive to patient values and preferences.

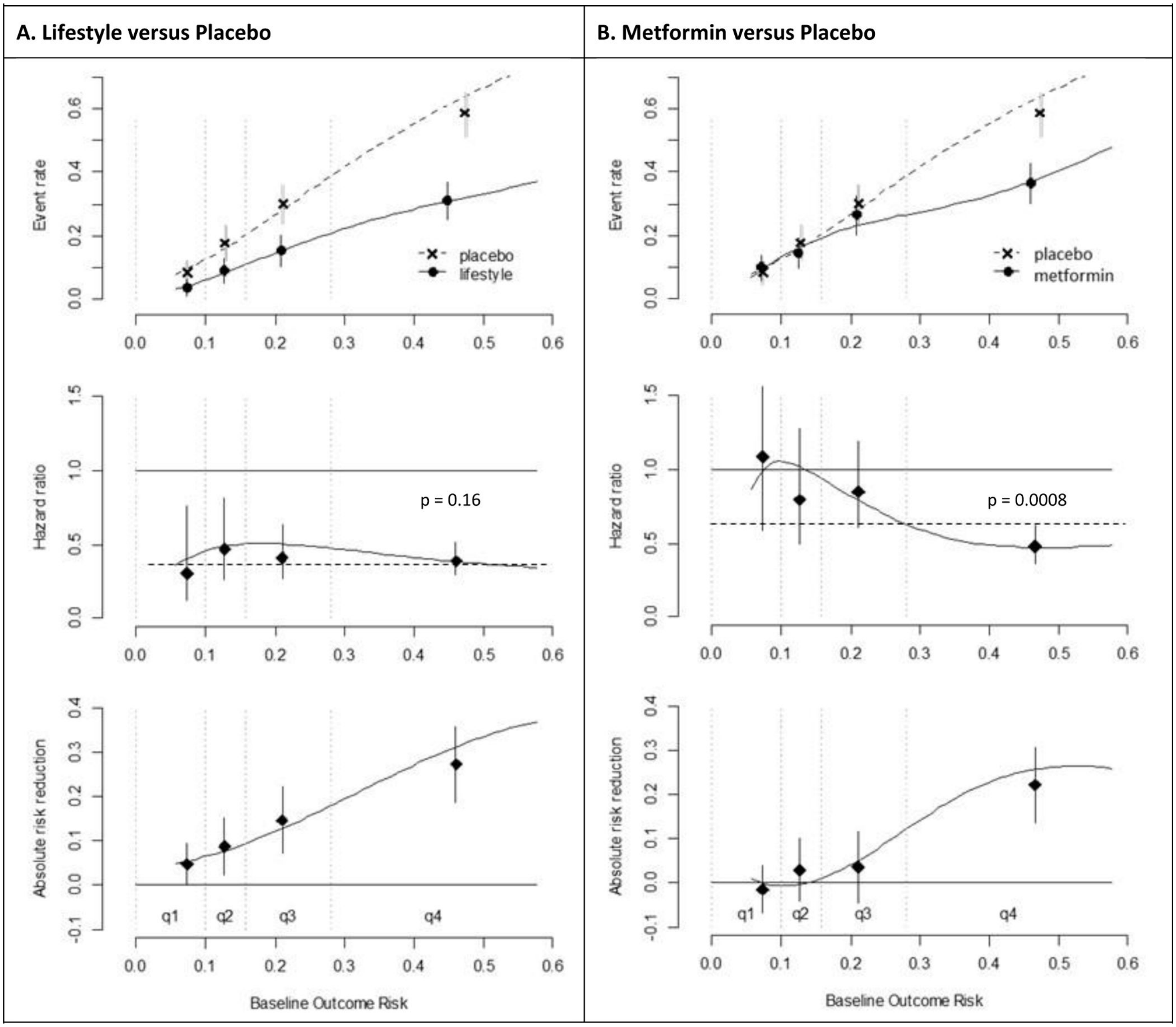


Figure 2. Effects of Lifestyle Modification and Metormin versus Usual Care in Patients with Prediabetes at Different Risks of Developing Diabetes[71]

Figure 2 presents HTE analysis of the Diabetes Prevention Program (DPP) Trial as a function of baseline risk[71]. Event rates (top graph), hazard ratios (middle graph) and absolute effects (lower graph) are shown. Both lifestyle modification (left panel) and metformin (right panel) are compared to usual care as a function of baseline risk. For lifestyle modification, a consistent 58% reduction in the hazard of developing diabetes over three years was found across all levels of risk. This consistent relative effect yields HTE on the absolute scale of potential clinical importance. In contrast, the effects of metformin are heterogeneous on the both the hazard ratio scale and on the absolute scale. Penalized splines were used to model the relationship between the linear predictor of risk and the time to event outcome. Vertical lines denote 95% confidence intervals and p-values are based on the null hypothesis of no effect modification tested using the linear predictor of risk in a Cox model.

The dashed lines show the average effects in the trial. Prediction of incient diabetes with an external model derived from the Framingham cohort yielded a similar pattern[123].

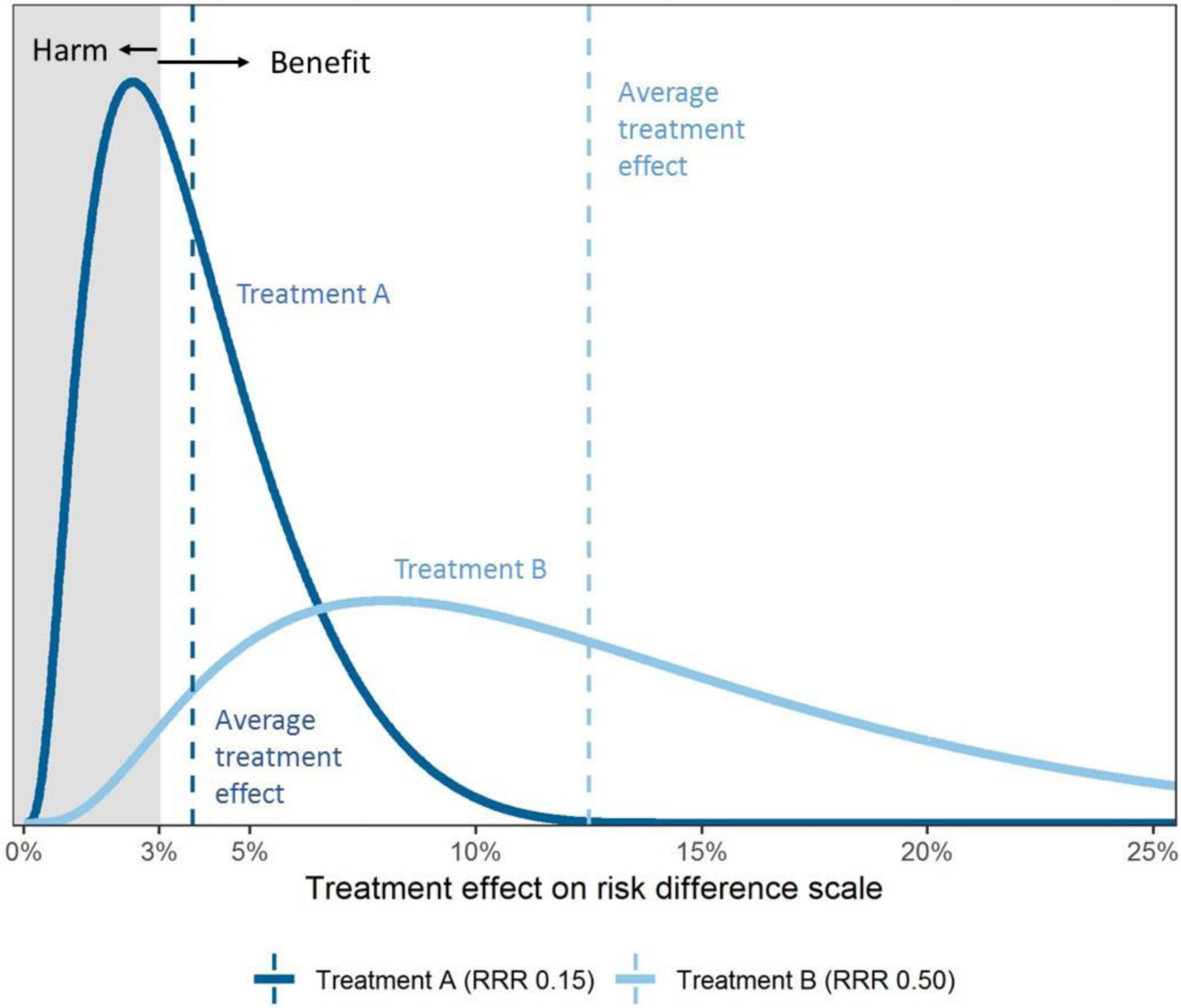
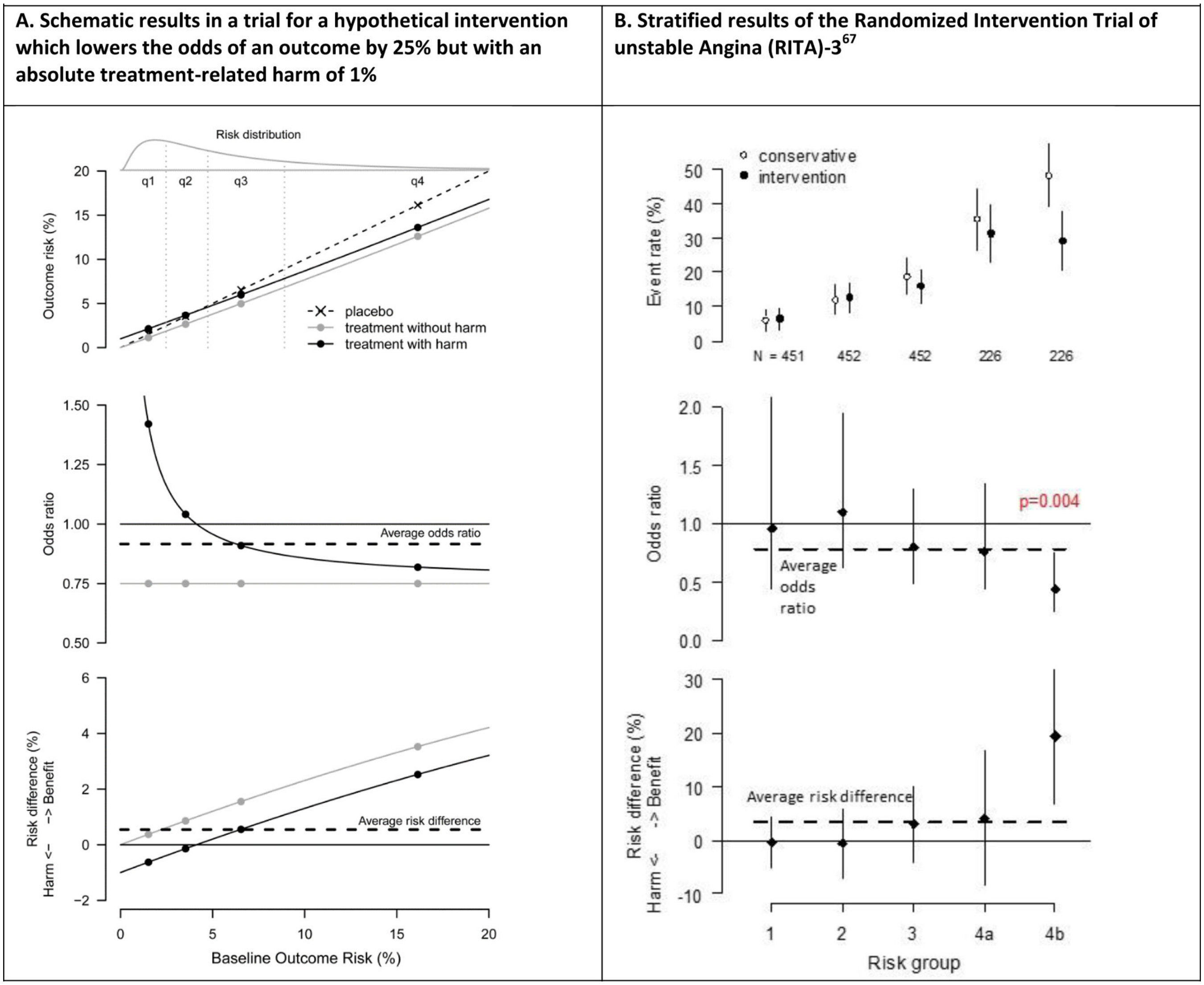


Figure 3: The Value of a Risk Modeling Approach is Likely to be Greater when the Average Treatment Effect in a Trial (Treatment A) is Near a Decision Threshold

Figure 3 depicts the anticipated influence of a risk modeling approach in two trials testing different treatments in the same population, one of a treatment (A) with a slightly favorable benefit-harm trade-off, and the other of a treatment with an extremely favorable benefit-harm trade-off (treatment B). Under both conditions, the control event rate is 25% and the minimal clinically significant difference (MCSD, i.e., the absolute benefit that would justify the experimental therapy) is 3 percentage points. (For simplicity, we display a single MCSD, with grey shading corresponding to portions of the population that should not be treated, but this value varies according to individual patient values and preferences.) A risk modeling approach would be of substantially greater value for the trial of therapy A, with the slightly favorable trade-off (with a relative risk reduction [RRR] of 0.15; absolute risk difference = 3.75%, just above the MCSD), compared to the trial of the therapy B with the extremely favorable trade-off (RRR of 0.5; risk difference = 12.5%, substantially above the MCSD).

The distributions show the anticipated risk differences that emerge with a constant RRR when the same moderately-predictive risk prediction model (i.e., with a c-statistic = ~0.70) is applied to the population. In the slightly favorable treatment condition (A), harms outweigh benefits in almost half the trial population (43%), despite overall results showing benefit on average. In the extremely favorable treatment condition (B), treatment remains worthwhile in virtually the entire population (97%). Thus, applying the risk modeling approach is very valuable in the low benefit condition, as it reclassifies many patients as treatment-unfavorable who would otherwise have been treated based on the overall result.



A. Schematic results in a trial for a hypothetical intervention which lowers the odds of an outcome by 25% but with an absolute treatment-related harm of 1%

B. Stratified results of the Randomized Intervention Trial of unstable Angina (RITA)-3[67]

Figure 4. Schematized and Actual Risk-based Heterogeneous Treatment Effects

This figure schematically depicts outcome risks for a trial testing a hypothetical intervention with an odds ratio of 0.75 but with an absolute treatment-related harm of 1% (shown in the top panel). Observed odds ratios (middle panel) and risk differences (bottom panel) are shown. Overall trial results are dependent on the average risk of the enrolled trial population. When the average risk is ~7% (as above), a well-powered study would detect a positive overall treatment benefit (shown by the horizontal dashed line in the middle and bottom panels). However, a prediction model with a C-statistic of 0.75 generates the risk distribution at the top of the figure. A treatment-by-risk interaction emerges (middle panel). Whether or not this interaction is statistically significant, examination of treatment effects on the absolute risk difference scale (bottom panel) reveals harm in the low risk group and very substantial benefit in the high risk group, both of which are obscured by the overall summary results. Conventional one-variable-at-a-time subgroup analyses are typically inadequate to disaggregate patients into groups that are sufficiently heterogeneous for risk such that benefit-harm trade-offs can misleadingly appear to be consistent across the trial population. Baseline risk is logit normal distributed with mu=−3 and sigma=1 (the log odds are normally distributed). Figure adapted from Kent DM et al. JAMA 2007.[3]

The RITA-3 trial (N=1810) tested early intervention versus conservative management of non-ST-elevation acute coronary syndrome. Results for the outcome of death or non-fatal myocardial infarction at 5 years are shown above, stratified into equal-sized risk quarters using an internally-derived risk model; the highest risk quarter is sub-stratified in halves (groups 4a and 4b). Event rates with 95% confidence intervals (top panel), odds ratios (middle panel), and risk difference (bottom panel) are displayed. The risk model is comprised of the following easily obtainable clinical characteristics: age, sex, diabetes, prior MI, smoking status, heart rate, ST depression, angina severity, left bundle branch block, and treatment strategy. As in the schematic diagram to the left, the average treatment effect seen in the summary results (horizontal dashed line in middle and bottom panels) closely reflect the effect in patients in risk quarter 3, while fully half of patients (q1 and q2) receive no treatment benefit from early intervention. Absolute benefit (bottom panel) in the primary outcome was very pronounced in the eighth of patients at highest risk (4b). A statistically significant risk-by-treatment interaction* can be seen when results are expressed in the odds ratio scale (middle panel). Such a pattern can emerge if early intervention is associated with some procedure-related risks that are evenly distributed over all risk groups, eroding benefit in low risk but not high risk patients, as illustrated schematically in Figure 4A.

*The interaction p value is from a likelihood ratio test for adding an interaction between the linear predictor of risk and treatment assignment (one degree of freedom).

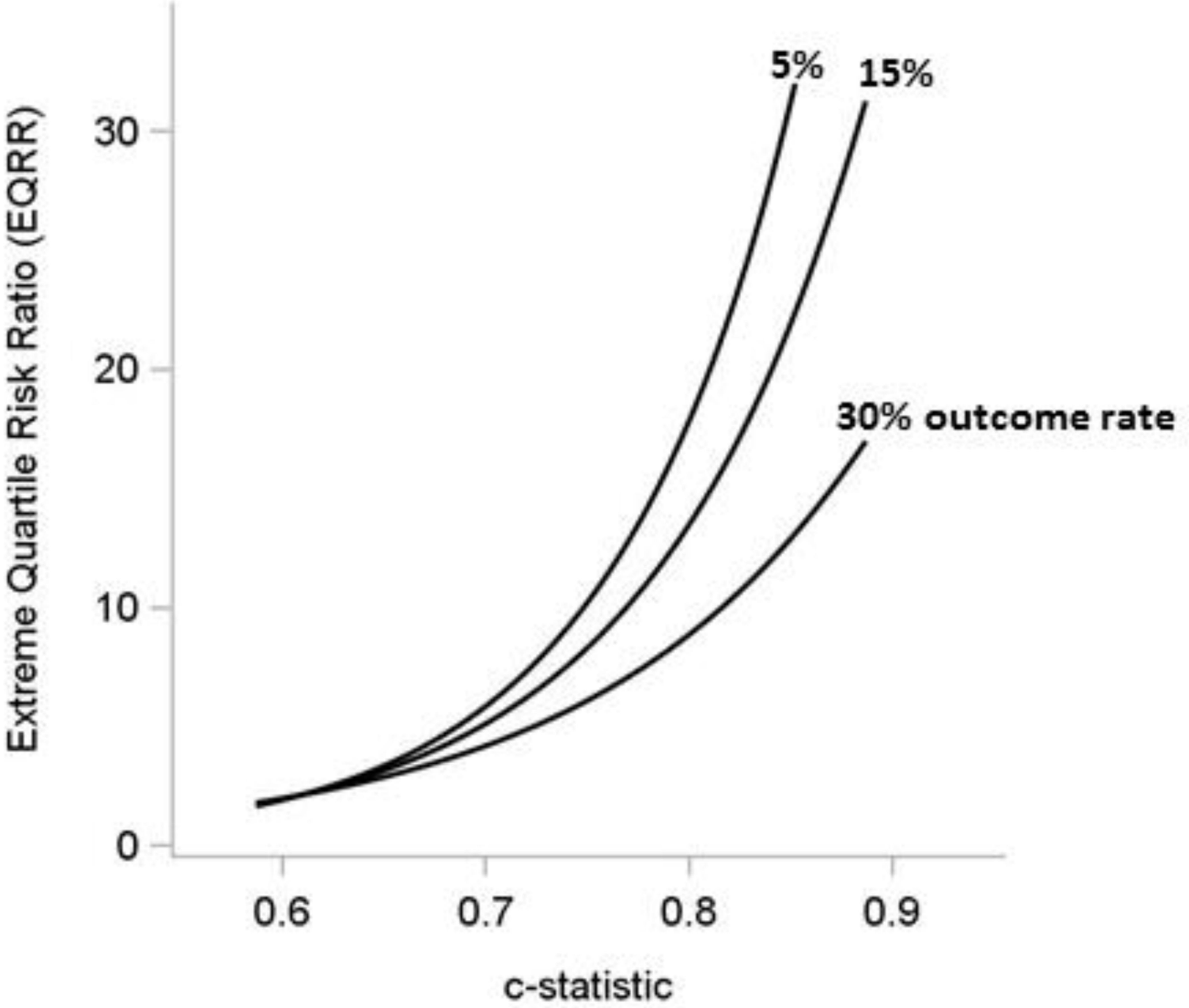


Figure 5: Risk heterogeneity increases with higher discrimination – Extreme Quartile Risk Ratio Increases With Increasing C-Statistic, Especially at Low Outcome Rates

The curves above depict the relationship between the c-statistic and extreme quartile risk ratio (EQRR) – that is, the risk in the highest quartile compared to the risk in the lowest quartile – for different outcome rates across 32 trials.[45] Unsurprisingly, the degree of risk heterogeneity (as represented by the EQRR) is strongly related to the discriminatory power of the prediction model. The relationship is strongest when the overall outcome rates are low. The c-statistic and EQRR both reflect how well the risk factors predict the outcome in a given population. For reference, in a trial with an outcome rate of 15%, a predictive model with a c-statistic of 0.80 is anticipated to yield an outcome rate that is 13-fold higher in the high risk quartile compared to the low risk quartile. When the outcome rate is lower (5%), this ratio is expected to be greater than 20-fold for a model with similar discrimination.

Patient groups with such different outcome risks are unlikely to have similar benefit-harm trade-offs for most therapies, even thought they may be included in the same trial.

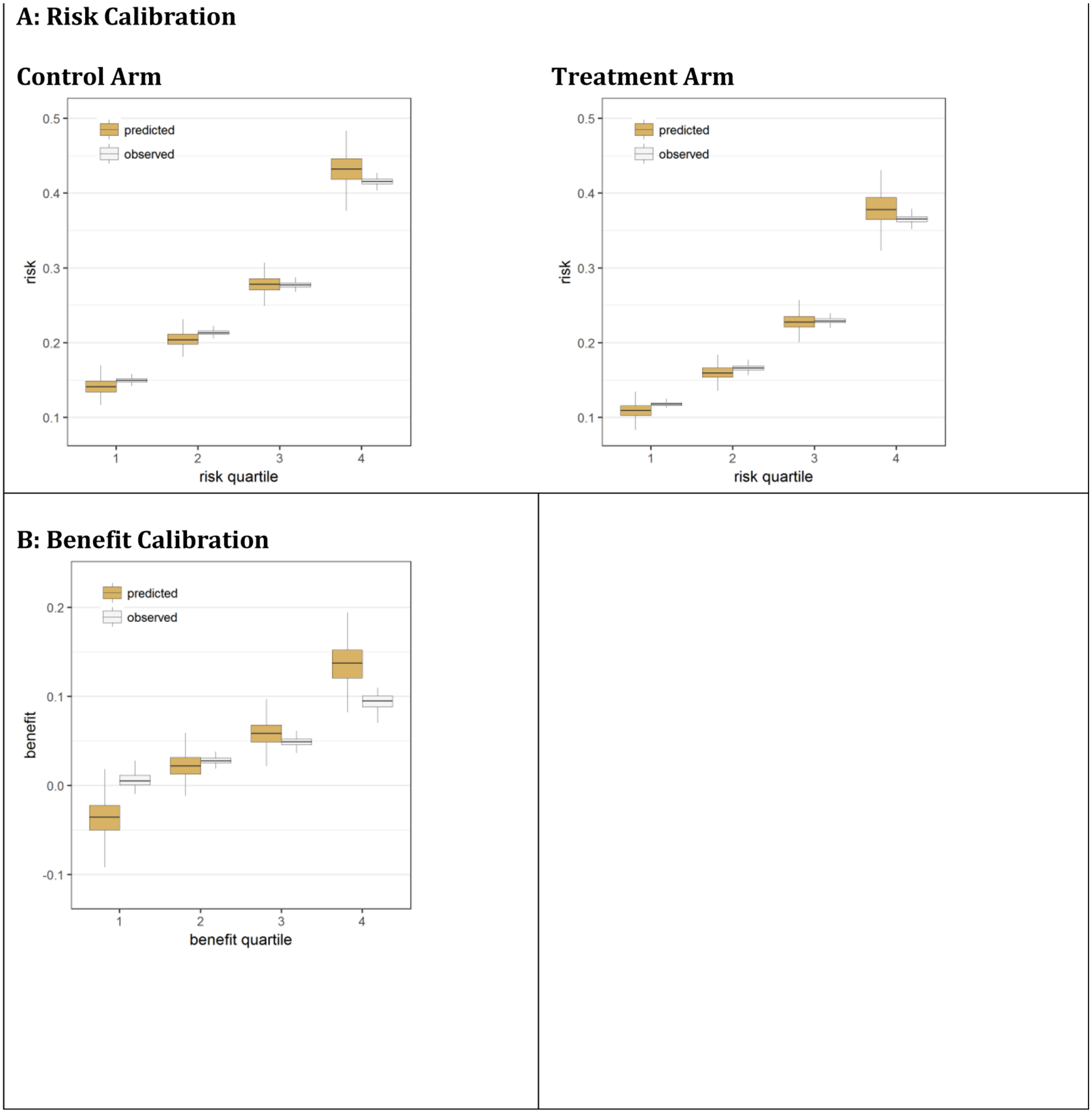


Figure 6: Evaluating Model Performance: A Comparison of Conventional Outcome Risk Calibration versus Treatment Effect (Benefit) Calibration

These data represent box plots of predicted and observed event rates by quartiles of predicted risk in the control and treatment arm of a hypothetical RCT (500 simulations; panel A). These rates appear to demonstrate appropriate model calibration. However, examining the same data for predicted and observed benefit (differences in event rates) by quarters of predicted benefit (panel B) reveals very poor model calibration at the extreme quarters. This poor calibration occurs because miscalibration for the risk difference includes error from both the control and treatment arm, and because the scale of the risk difference is much smaller than that for the outcome risk. These data was generated from a simulation of a prediction model that included 12 treatment effect interactions, 6 of which represented true interactions. The boxes represent, in line with Tukey's definition, the 25% quantile to the 75% quantile (with the median shown). The lower and upper whiskers include the most extreme observations within the range of 1.5 times the interquartile range, from the 25% and 75% quantiles, respectively.

Table 1:

Treatment Effect is Mathematically Dependent on the Control Event Rate

Measure	**Definition**
Absolute Risk Difference	**CER**-TER
Relative Risk Reduction	$1-\frac{\mathrm{TER}}{\mathbf{CER}}$
Odds Ratio	$\frac{\mathrm{TER/(1\text{-}TER)}}{\mathbf{CER/(1\text{-}CER)}}$

CER=control event rate

TER=treatment event rate

Table 2:

Hypothetical Example Presentation of the Effects of Model-based Decision Making

Strategy	Number of Patients Treated	Number of Events	Decrease in Event Rate
Treat no patient	0	250	--
Treat all patients	1000	200	50
Treat only those with a predicted benefit >5%	400	215	35

Table 3.

Methodological Literature on the Conduct of Regression Modeling Approaches to Treatment Effect Heterogeneity Analysis

Approach	Description
Risk-modeling approaches[7;45;62;98;117;118;120;183–186]	Using a multivariable risk model developed blinded to treatment effect, analyze the relationship between baseline risk and treatment effect on the relative and on the absolute scale. While treatment effect modification on the relative scale across different levels of baseline risk is considered, treatment effect modification on the relative scale for individual risk factors is not.
Treatment effect modeling approaches[120;143-145;151;186–193] Subgroup identification (2-step process)[143;144;187] Individualized treatment effects (1-step process)[120;145;186;188–190]	Use both the main effects of risk factors and interaction effects with treatment assignment (on the relative scale) to estimate more individualized treatment effects. They can be used either for defining patient subgroups with similar expected treatment effects or for predicting individualized treatment effects on the absolute scale for future patients.
Optimal treatment regimens[132–142]	Classify patients into those who benefit from treatment (positive individualized treatment effect) and those who do not (negative individualized treatment effect), through the identification of modifiers of treatment effects on the relative scale.

Table 4.

A Meta-Research Agenda for Predictive Approaches to Treatment Effect Heterogeneity *More research is needed to:*

High Priority Research Needs
Better understand the value of HTE methods through empirical analyses across a wider range of clinical domains.
Determine optimal approaches to penalization/regularization in effect modeling to mitigate the risks of overfitting; or other methods that permit the exploration of plausible hypotheses of effect modification on the relative scale while strongly protecting against false positive findings.
Determine optimal methods to simultaneously predict multiple risk dimensions (e.g. risk of the primary outcome versus risk of treatment-related harm) and/or optimal approaches to combine models predicting these outcome risks for improved benefit-harm discrimination.
Determine the optimal measures to evaluate models intended to predict treatment benefit.
Identify heuristics or general principles to judge the adequacy of sample sizes for predictive analytic approaches to HTE, particularly for treatment effect modeling.
Determine optimal methods to validate, recalibrate, and/or update models predicting treatment effect in the absence of new randomized trials.
Other Research Needs
Determine optimal methods to combine predictive HTE analyses with methods that permit the estimation of direct treatment effects or adherence-adjusted effects in the presence of drop out, loss to follow up, poor adherence and treatment switching.
Identify the appropriate clinical contexts for which modeling multiple dimensions of risks (e.g., risk of the primary outcome; risk of treatment-related harm; risk of an important competing outcome) is important and feasible for adequate disaggregation of benefit-harm trade-offs.
Determine methods to analyze trials with multiple important outcomes, or outcomes where differences in treatment effect may be related to the choice of follow-up time.
Better understand the impact of different missingness mechanisms and develop principled methods for dealing with missing data in the context of subgroup identification.
Determine methods to permit models predicting treatment effect to cope with missing data in clinical practice.
Determine optimal methods for modeling the functional form of the risk-by-treatment interaction to translate risk stratified results from trials into continuous treatment effect predictions for clinical application.
Address concerns about differential fit between arms from endogenously derived risk models when randomization is unbalanced (e.g., 1:2 or 1:3 randomization).
Determine optimal methods to achieve balance in covariates across subgroups in observational databases.
Determine whether novel methods, including machine learning techniques, have distinct advantages over traditional statistical approaches for predicting treatment benefit.
Examine how to best extend these approaches to other trial designs (e.g. longitudinal studies, dynamic treatment regimens)