The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration

David M. Kent, MD, MS; David van Klaveren, PhD; Jessica K. Paulus, ScD; Ralph D'Agostino, PhD; Steve Goodman, MD, MHS, PhD; Rodney Hayward, MD; John P.A. Ioannidis, MD, DSc; Bray Patrick-Lake, MFS; Sally Morton, PhD; Michael Pencina, PhD; Gowri Raman, MBBS, MS; Joseph S. Ross, MD, MHS; Harry P. Selker, MD, MSPH; Ravi Varadhan, PhD; Andrew Vickers, PhD; John B. Wong, MD; and Ewout W. Steyerberg, PhD

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

Ann Intern Med. 2020;172:W1-W25. doi:10.7326/M18-3668 **Annals.org**
For author affiliations, see end of text.
This article was published at Annals.org on 12 November 2019.

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 is mounting recognition within evidence-based medicine 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, interest is growing in understanding how a treatment's effect can vary across patients—a concept described as heterogeneity of treatment effects (HTE) (5–11).

Much literature exists on the limitations of conventional "1-variable-at-a-time" subgroup analyses, which serially divide the trial population into groups (for example, male vs. female or old vs. young) and examine the contrast in the treatment effect across these groups (12–22). The limitations include risks for 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 are also incongruent with the way clinical decision making occurs at the level of the individual patient, because patients have multiple attributes simultaneously that can affect the tradeoffs between the benefits and harms of the intervention. Individual patients thus belong to multiple subgroups, each of which may yield a different estimate of the treatment effect (4, 10).

The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement offers guidance relevant for "predictive" approaches to HTE analysis (26) that are designed to address some of the limitations mentioned in the previous paragraph. The goal of predictive HTE analysis is to provide individualized predictions of treatment effect, specifically defined by the difference between expected potential outcomes of interest with one intervention versus an alternative (4, 8). We refer to this as the "individualized treatment effect." We avoid the term "individual treatment effects" because this latter term confusingly suggests that treatment effects can be estimated at the person level; such effects are inherently unobservable in parallel-group clinical trials because only 1 of 2 counterfactual potential outcomes can be observed (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 (that is, 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 decision making for individual patients (29). By accounting for multiple variables simultaneously, predictive HTE analysis is foundational to the concept of personalization in evidence-based medicine (4).

DISTINCT APPROACHES TO PATH

The PATH Statement (26) 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 PATH

See also:

Related article XX
Editorial comment 1

Glossary

Effect modification: Occurs when the magnitude of the effect of the primary treatment or exposure on an outcome differs depending on the level of a third variable (e.g., patient characteristics). In the presence of effect modification, the use of an overall effect estimate is inappropriate.

Heterogeneity of treatment effects (HTE): Nonrandom variation in the direction or magnitude of a treatment effect, measured using clinical outcomes. HTE is fundamentally a scale-dependent concept, and therefore, for clarity, the scale should generally be specified.

Clinically important HTE: Occurs when variation in the risk difference across patient subgroups spans an important decision threshold, which depends on treatment burden (including treatment-related harms and costs). It is generally assessed on the absolute scale.

Statistically significant HTE: Occurs when the effect of an intervention is found to differ across levels of a covariate on a selected scale as evaluated against a null hypothesis statistical test. In regression analysis, HTE is typically tested on a relative scale (e.g., odds ratio or hazard ratio) by examining the statistical significance of an interaction term between a covariate and treatment.

Causal interaction: Occurs when manipulating a covariate would change the effect of the primary treatment on the outcome. Statistically significant covariate-by-treatment interaction does not necessarily imply causal interaction.

Predictive HTE analysis: The main goal of predictive HTE analysis is to develop models that can be used to predict which of 2 or more treatments will be better for a particular individual.

Risk-modeling approach: An approach to predictive HTE analysis where a multivariable model that predicts the risk for an outcome (usually the primary study outcome) is applied to disaggregate patients in trials to examine risk-based variation in treatment effects.

External models vs. endogenous or internally derived models: An external risk model has been developed from an external trial or cohort population that can be applied for HTE analysis of the trial. An endogenous or "internal" risk model is one developed directly on the trial population that does not include a term for treatment assignment.

Treatment effect–modeling approach: An approach to predictive HTE analysis that develops a model directly on randomized trial data to predict treatment effects (i.e., the contrast in outcome risks under 2 alternative treatment conditions). Unlike in risk modeling, the model incorporates a term for treatment assignment and permits the inclusion of treatment-by-covariate interaction terms.

Net benefit: A decision analytic measure that puts benefits and harms on the same scale. This is achieved by specifying an exchange rate based on the relative value of benefits and harms associated with interventions. The exchange rate is related to the probability threshold to determine whether a patient is classified as positive or negative for a model outcome, or (when applied to trial analysis) as treatment-favorable vs. treatment-unfavorable.

Overfitting: A key threat to the validity of a model where predictions do not generalize to new persons outside the sample under study. Overfitting occurs when a model conforms too closely to the idiosyncrasies or "noise" of the limited data sample on which it is derived.

Penalized regression: A set of regression methods, developed to prevent overfitting, in which the coefficients assigned to covariates are penalized for model complexity. Penalized regression is sometimes referred to as shrinkage or regularization. Examples of penalized regression include LASSO (least absolute shrinkage and selection operator) regression, ridge regression, and elastic net regularization.

Reference class: A group of similar case patients that is used to make predictions for an individual case patient of interest. The "reference class problem" refers to the fact that similarity can be defined in an indefinite number of ways.

Subgroup analysis: An analysis that examines whether specific patient characteristics modify the effects of treatment on an outcome.

Testimation bias (also known as the "winner's curse"): Refers to the fact that the effect sizes of newly discovered true (non-null) associations are inherently inflated on average. Testimation bias arises because of the use of thresholds in the process of discovering associations. Inflation is usually expected when both 1) an association has to pass a certain threshold of statistical significance to be deemed positive and 2) the study that leads to the discovery has suboptimal power.

Statement guidance focuses on identifying "clinically important HTE" (4, 7, 10), or variation in the risk difference across patient subgroups that may be sufficient to span important decision thresholds that reflect treatment-related harms and burdens. The statement offers guidance on 2 distinct approaches to predictive HTE analysis (4). With a "risk-modeling" approach, a multivariable model that predicts risk for an outcome (usually the primary study outcome) is first 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 approach, "effect modeling," 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 they are especially vulnerable to overfitting and false discovery of promising subgroup effects (or they require very large databases that are well powered for the detection of interaction effects) (30). Both approaches can be used to predict individualized treatment effects—that is, the *difference* in expected outcome risks under 2 alternative treatments, conditional on important clinical variables. A fuller introduction to risk and effect modeling is presented in prior literature (4).

In this PATH Statement explanation and elaboration, we expand on the intent and motivation (and reservations) regarding the statements, criteria, considerations, and caveats. Recommendations are explained in more detail and accompanied by clinical applications of selected methods and supporting methodological evidence where relevant. The **Glossary** defines terms relevant to these methods.

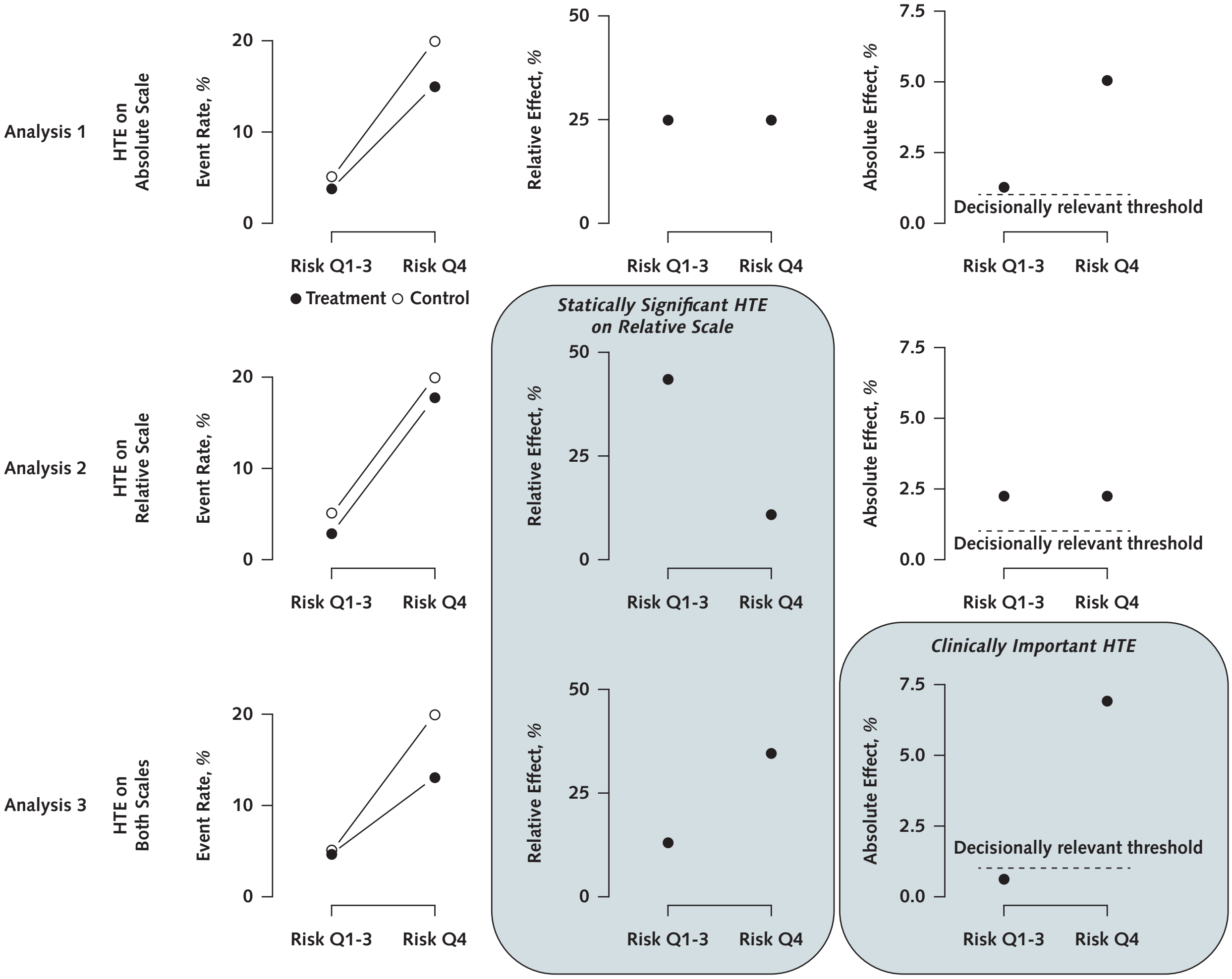
Clarification of Terms and PATH Statement Scope

The term "heterogeneous treatment effects" has been used in the literature in different ways. In this article, we define HTE as nonrandom variation in treatment effects across levels of a covariate (that is, a patient attribute or a score comprising multiple attributes), as measured on a selected scale, against a clinical outcome. It corresponds to the epidemiologic concept of effect measure modification but applies specifically to treatment effects. In clinical trials, HTE is identified by contrasting treatment effects on a chosen scale between subgroups and testing for statistical interactions. Of note, HTE, effect measure modification, and statistical interaction are all "scale-dependent" concepts—that is, their presence or absence depends on what scale is selected to measure treatment effect (31). The scale dependence of HTE is illustrated in **Figure 1**, which contrasts 3 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 DPP (Diabetes Prevention Program) trial, which tested lifestyle modification and metformin against usual care for prevention of diabetes (Figure 2). Although only 1 tested therapy showed statistically significant HTE on the relative (hazard ratio) scale, both therapies had substantial HTE on the clinically important absolute scale. The Glossary defines key terms described here and elsewhere in the PATH Statement and explanation and elaboration document (26).

HTE Analysis for Causal Interaction Versus for Prediction and Decision Making

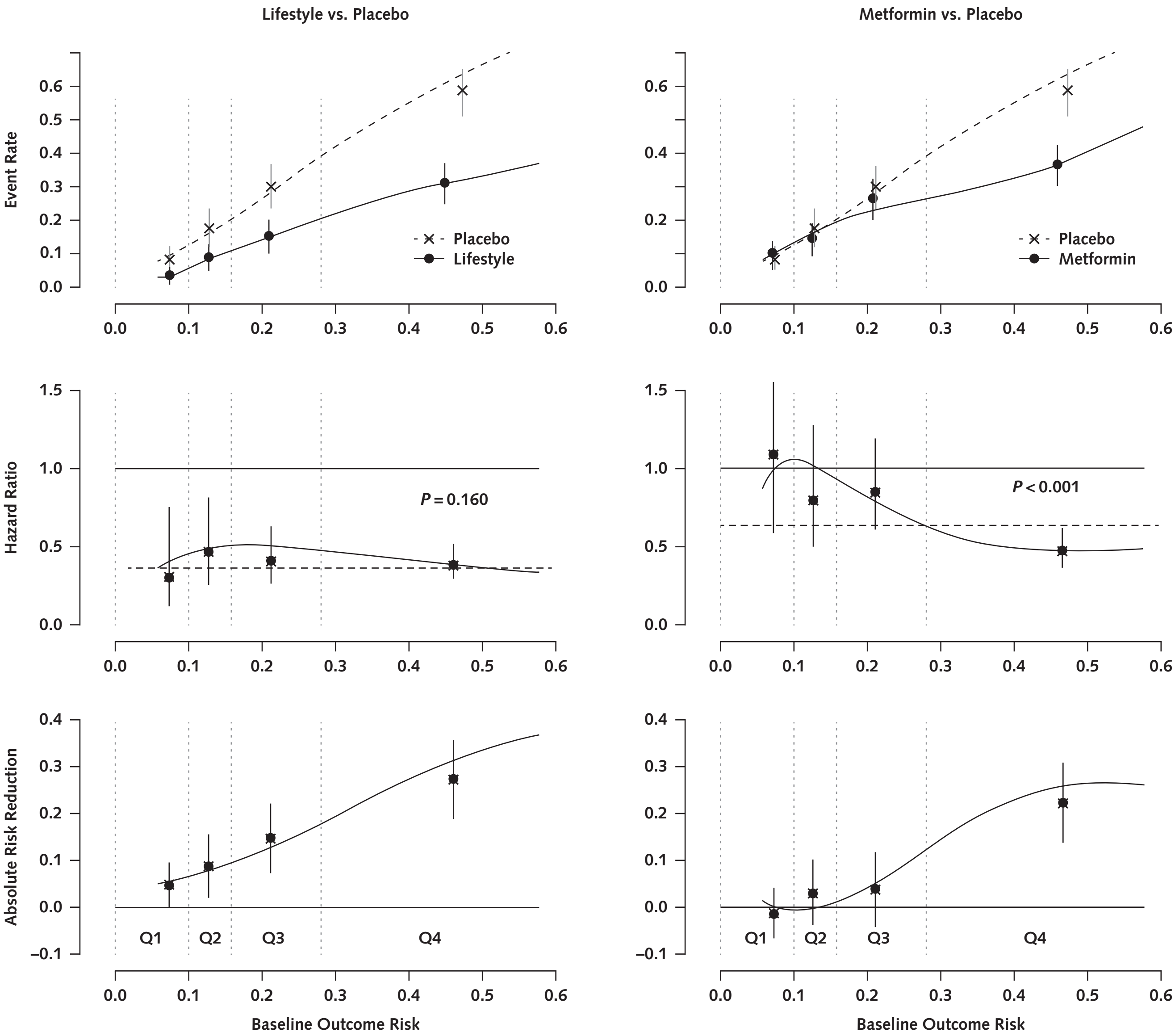
We also note that HTE (and statistical interactions) are used to make 2 very different kinds of inferences: causal inferences (for example, regarding causal or biological interaction) and inferences for clinical decision making. Although the importance of statistical interac-

Figure 1. The scale dependence of HTE.



All 3 scenarios are drawn from hypothetical trials with the same overall results (outcome rate, 8.8% in the control group [*open circles*] vs. 6.6% in the treatment group [*closed circles*]) and depict outcomes in low-risk groups (75% of patients, Q1-3) and high-risk groups (25% of patients, Q4) (where control event rates are 5% and 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 the third row, effect heterogeneity is present on both the relative and the absolute scale. The statistical significance of HTE is typically tested on the relative scale (*middle column*) because 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, although HTE is present 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. Of note, the presence of statistically significant interaction (on the relative scale) does not imply clinically important HTE, and the absence of 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 vs. another. Establishing causal interaction effects is 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. HTE = heterogeneity of treatment effects; Q1 = first risk quarter (lowest); Q2 = second risk quarter; Q3 = third risk quarter; Q4 = fourth risk quarter (highest).

Figure 2. Effects of lifestyle modification and metformin vs. usual care in patients with prediabetes at different risks for diabetes.



This figure presents HTE analysis of the DPP (Diabetes Prevention Program) trial as a function of baseline risk (32). Event rates (*top*), hazard ratios (*middle*), and absolute effects (*bottom*) are shown. Both lifestyle modification (*left*) and metformin (*right*) are compared with usual care as a function of baseline risk. For lifestyle modification, a consistent 58% reduction in the hazard of developing diabetes over 3 y 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 both the hazard ratio scale and 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% CIs, and *P* values are based on the null hypothesis of no effect modification tested using the linear predictor of risk in a Cox model. In the hazard ratio graphs, the dashed lines show the average effects in the trial and the horizontal lines at 1.0 refer to the null effect on this scale. The horizontal lines at 0 in the absolute risk reduction graphs refer to the null effect on this scale. Prediction of incident diabetes with an external model derived from the Framingham cohort yielded a similar pattern (33). HTE = heterogeneity of treatment effect; Q1 = first risk quarter (lowest); Q2 = second risk quarter; Q3 = third risk quarter; Q4 = fourth risk quarter (highest).

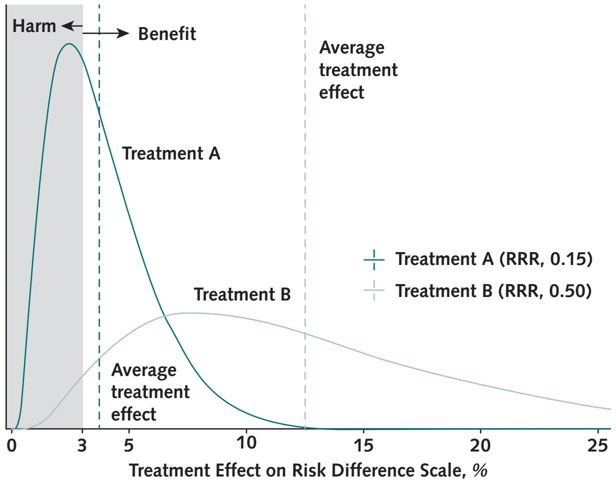
tions is often stressed for HTE analysis, these inferences are only weakly related to the presence of "statistically significant HTE." Statistical interactions should not be confused with causal interactions, and statistically significant HTE should also not be conflated with clinically important HTE. These issues are described briefly in the following paragraphs.

In regression models examining HTE, *causal inferences* depend on interpretation of model *inputs* (that is, 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. Many methodologists believe that interaction on a multiplicative (relative) scale is stronger evidence in support of a causal interaction than interaction on an absolute scale (although this is by no means a universal view) (34–40). 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) because 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, we would need to control for all relevant differences in these other variables (that is, 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.

In regression models examining HTE, *inferences for clinical decision making* depend on interpretation of model *outputs*. Because of this, such analyses have been called "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 2 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. **Figure 1** illustrates the scale dependence of effect heterogeneity. We also note that controlling for confounding factors (that is, factors that differ between levels of the subgrouping variable) is not necessary for prediction (35, 41).

A new term, "risk magnification," has recently been coined to describe a method of identifying high-risk, high-benefit patients (42, 43). 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 (44–46). Risk magnification is distinct from the risk-modeling approach described here because it can be applied without any data based on the assumption of a constant relative treatment effect. Indeed, the use of the pooled cohort equations (47) (also known as the atherosclerotic cardiovascular disease risk estimator) to target statin therapy to patients at high risk for coronary heart disease might be described as an application of "risk magnification." Because many (observed and unobserved) patient attributes change across 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. In addition, the rate of adverse events may also differ across levels of baseline risk (recommendation 9 in **Figure 3** of the PATH Statement [26]). Examining randomized data stratified by a risk model also permits these other (nonprimary) outcomes to be examined across levels of risk.

Figure 3. Value of a risk-modeling approach when the average treatment effect in a trial (treatment A) is near a decision threshold.



This figure depicts the anticipated influence of a risk-modeling approach in 2 trials testing different treatments in the same population, one (treatment A) with a slightly favorable benefit–harm tradeoff and the other (treatment B) with an extremely favorable benefit–harm tradeoff. Under both conditions, the control event rate is 25% and the MCSD (i.e., the absolute benefit that would justify the experimental therapy) is 3 percentage points. (For simplicity, we show a single MCSD, with gray 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 treatment A, with the slightly favorable tradeoff (RRR, 0.15; absolute risk difference, 3.75% [just above the MCSD]), than for the trial of treatment B, with the extremely favorable tradeoff (RRR, 0.50; 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 of about 0.70) is applied to the population. In the slightly favorable treatment condition (A), harms outweigh benefits in almost half of the trial population (43%) despite overall results showing benefit on average. In the extremely favorable treatment condition (B), treatment remains worthwhile in almost the entire population (97%). Thus, applying the risk-modeling approach is very valuable in the low-benefit condition because it reclassifies many patients as treatment-unfavorable who would otherwise have been treated on the basis of the overall result. MCSD = minimal clinically significant difference; RRR = relative risk reduction.

PATH Statement Criteria for When Risk Modeling Is Likely to Be of Value

Included Criteria

The following criteria identify the features of the data, modeling, and clinical decisions that are common to scenarios in which the application of predictive HTE analyses to treatment comparisons is likely to be relevant for individualized clinical decision making (**Figure 2** of in the PATH Statement [26]). The motivation and reservations regarding these criteria are elaborated.

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

When clinical trials are null, researchers may be tempted to find subgroups of patients in whom the treatment might work. However, clinically important subgroup effects discovered through risk modeling 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 (48). More "aggressive" effect-modeling approaches (that is, those reliant on including treatment-by-covariate interaction terms within a prediction model) may identify groups of patients who seem to benefit, but such approaches are also likely to yield spurious false-positive results (28, 49). Such "treatment-favorable" groups may be suggested by pure chance and are aggravated by multiple testing, even when treatments have no effect whatsoever (30). Thus, predictive HTE analyses are more appropriately done on interventions for which an overall effect has been established. Despite a tendency to focus on any subgroup with a positive effect 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 group and might nullify an overall effect. For example, angiotensin-converting enzyme inhibitors both cause and prevent renal insufficiency, thrombolytic therapy both causes (via hemorrhage) and prevents (via reperfusion) functional disability in patients presenting with acute stroke (50), carotid surgery can both cause and prevent ischemic stroke in patients with carotid stenosis (51), and antiarrhythmic agents both cause and prevent serious cardiac arrhythmias (52, 53). 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, through application of either an outcome risk model or a model identifying patients at high risk for treatment-related harm (that is, for "treatment deselection") (50, 54, 55). Such an approach is possible particularly when the predictors of outcome risk are poorly (or negatively) correlated with predictors of risk for treatment-related harm. For example, excluding patients at high risk for thrombolytic-related intracranial hemorrhage in stroke (such as older patients with higher blood pressure) may uncover benefit in patients at lower risk (50).

2. When the benefits and harms or burdens of a given intervention are finely balanced (that is, 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 risk in the population (56, 57). The threshold depends on clinical context and involves weighing the expected benefits against the expected harms or costs of a decision. This assessment can be done with formal decision analysis but is more often done informally.

The idea 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 this 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 by considering the benefits that accrue to those switching (58–60). Beyond measuring model accuracy, various methods have been proposed to evaluate the potential effect of model use on a particular decision, including risk stratification tables (61), relative utility curves (62), predictiveness curves (63), and decision curves (59).

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 (that is, when *net* benefit is near 0), any additional prognostic or predictive information is likely to be especially useful for determining the better therapy for a particular patient.

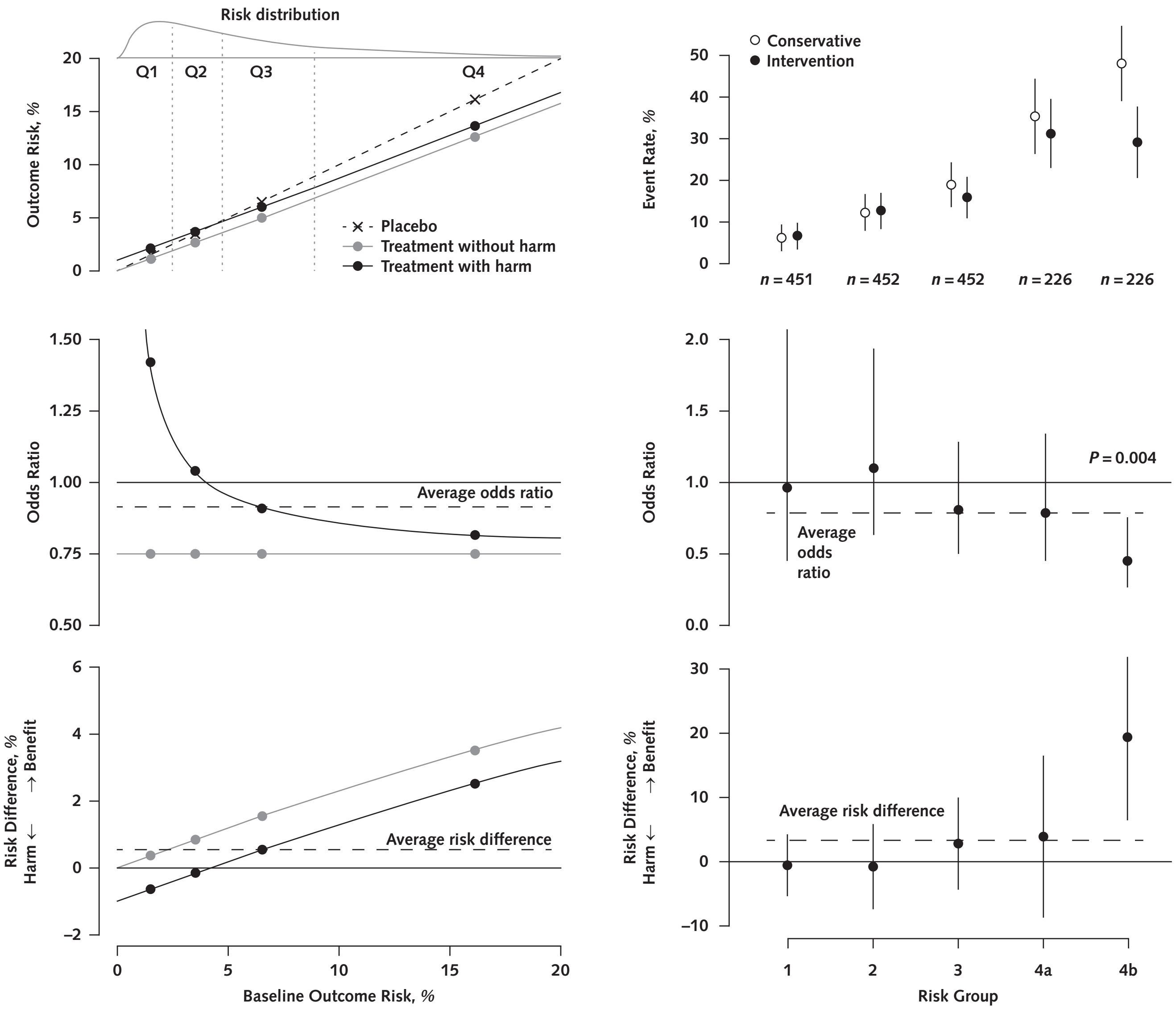
The relative utility of risk prediction when average risk is near versus far from a threshold is described schematically in **Figure 3**, which shows the distribution of expected benefits when a risk model with a c-statistic of 0.70 is applied to a population with an average risk of 25%. We consider the following 2 treatment conditions: 1) a treatment with a moderate average treatment effect (relative risk reduction, 15%; risk difference, 3.8%) and an average benefit–harm tradeoff that is slightly favorable (vs. a minimal clinically important difference of 3%) and 2) a treatment with a large treatment effect (relative risk reduction, 50%; risk difference, 12.5%) and an average benefit–harm tradeoff that is clearly favorable. All else being equal, the net benefit of risk modeling would be much greater when benefit–harm tradeoffs are more finely balanced because risk stratification would reveal many patients (almost half of 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 (**Figure 4**, *left*) 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 because 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 $CHADS_2$ score (and its variants) is used to target anticoagulation to patients with nonvalvular atrial fibrillation, but patients with higher $CHADS_2$ scores are also known to be at higher risk for anticoagulation-related hemorrhage (65). Given the potential (positive or negative) correlation between benefits and harms, we recommend that harms be reported in each risk stratum to support stratum-specific evaluation of benefit–harm tradeoffs (recommendation 9 in **Figure 3** of the PATH Statement [26]).

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

This criterion is related to the previous 2, in that HTE will be most important to decision making in the

Figure 4. Schematized (*left*) and actual (*right*) risk-based heterogeneous treatment effects.



Q1 = first risk quarter (lowest); Q2 = second risk quarter; Q3 = third risk quarter; Q4 = fourth risk quarter (highest). **A.** Schematic results in a trial for a hypothetical intervention that lowers the odds of an outcome by 25% (odds ratio, 0.75) but has an absolute treatment-related harm of 1%. Outcome risks (*top*), observed odds ratios (*middle*), and risk differences (*bottom*) are shown. Overall trial results are dependent on the average risk for the enrolled trial population. When the average risk is about 7% (as in this example), 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 in the top panel of the figure. A treatment-by-risk interaction emerges (*middle*). Regardless of whether this interaction is statistically significant, examination of treatment effects on the absolute risk difference scale (*bottom*) shows 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 "1-variable-at-a-time" subgroup analyses are typically inadequate to disaggregate patients into groups that are sufficiently heterogeneous for risk, so benefit–harm tradeoffs can misleadingly seem to be consistent across the trial population. Although this figure shows idealized relationships between risk and treatment effects, these relationships will be sensitive to how risk is described (i.e., what variables are in the risk model). Baseline risk has a logit-normal distribution, with $\mu = -3$ and $\sigma = 1$ (the log odds are normally distributed). Adapted from reference 3.

Q1 = first risk quarter (lowest); Q2 = second risk quarter; Q3 = third risk quarter; Q4 = fourth risk quarter (highest). **B.** Stratified results of RITA-3 (Randomized Intervention Trial of unstable Angina 3) (64). The RITA-3 trial (*n* = 1810) tested early intervention vs. conservative management of non–ST-segment elevation acute coronary syndrome. Results for the outcome of death or nonfatal myocardial infarction at 5 y are shown, stratified into equal-sized risk quarters using an internally derived risk model; the highest-risk quarter is substratified into halves (groups 4a and 4b). Event rates with 95% CIs (*top*), odds ratios (*middle*), and risk differences (*bottom*) are shown. The risk model comprises the following easily obtainable clinical characteristics: age, sex, diabetes, prior myocardial infarction, smoking status, heart rate, ST-segment 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 reflects the effect in patients in risk group 3, whereas half of patients (risk groups 1 and 2) receive no treatment benefit from early intervention. Absolute benefit (*bottom*) in the primary outcome was very pronounced in the eighth of patients at highest risk (risk group 4b). A statistically significant risk-by-treatment interaction can be seen when results are expressed in the odds ratio scale (*middle*) (the interaction *P* value is from a likelihood ratio test for adding an interaction between the linear predictor of risk and treatment assignment). 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 the left panel.

presence of a *qualitative* interaction—meaning that some patients benefit while others are harmed. By definition, qualitative interactions do not arise where treatments are 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 shown in simulation studies (49, 66) and observed empirically for carotid endarterectomy (51), stroke prevention in nonvalvular atrial fibrillation (67, 68), and medical or mechanical reperfusion in ST-segment elevation myocardial infarction (64, 69, 70). Treatment-related harm may be reflected in the primary outcome or ascertained as a separate outcome (such as acute kidney injury, major hemorrhage [54, 71], or serious bone fractures) (55, 72). Risk modeling may also be appropriate for particularly burdensome interventions (for example, major lifestyle commitments [32, 73] and treatment-related costs) (74, 75).

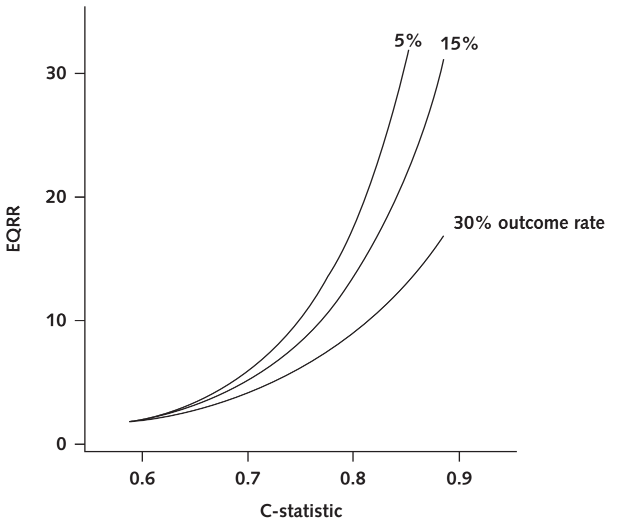
4. When several large, well-conducted RCTs of contemporary interventions are available and appropriate for pooling in individual patient meta-analysis.

Clinical trials are typically powered to detect an overall average treatment effect in a population, not to 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 required to detect a subgroup effect is at least 4-fold larger than for a main effect, even under favorable conditions (that is, well-balanced subgroups and overall and subgroup effects that are similar in size), but will generally be much larger than that (23, 24). Although 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 tradeoffs across strata. We also emphasize the need for contemporary trials because they are more likely to be relevant for contemporary clinical care. In addition, although such analyses have not been consistently successful in discovering reliable treatment effect interactions (76), combining data is likely to lead to substantial increases in risk heterogeneity in the study population (criterion 5), increasing the likelihood of uncovering clinically important HTE (that is, on the risk difference scale) (77, 78).

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

Risk heterogeneity depends on the presence of significant factors to predict outcome risk and differences in the distribution of these factors across the population (that is, a nonhomogeneous population). In the absence of factors that can predict outcome risk, no risk heterogeneity exists. Conversely, risk heterogeneity is highest in the presence of good discrimination (that is, a high c-statistic). Indeed, risk heterogeneity in a given population is a model-dependent property (79). **Figure 5** shows the empirical relationship between the c-statistic and the extreme-quartile risk ratio (the ratio of outcome rate in the highest risk quartile to that in the lowest risk quartile [80]) across 32 publicly available trials (48). Because clinical prediction models are abundant (7, 81, 82), 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 than trials with more narrowly restrictive enrollment criteria. Nevertheless, risk heterogeneity seems to be substantial even in classic efficacy trials (32, 48, 64, 70, 71). Because individual patient meta-analyses (that is, combining trials) have even higher risk heterogeneity at the patient level, they are an ideal substrate for these analyses (77, 78).

Figure 5. Risk heterogeneity increases with higher discrimination, and EQRR increases with increasing c-statistic, especially at low outcome rates.



The curves depict the relationship between the c-statistic and EQRR—that is, the risk in the highest quartile compared with the risk in the lowest quartile—for different outcome rates across 32 trials (46). 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 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 highest risk quartile than in the lowest risk quartile. When the outcome rate is lower (5%), this ratio is expected to be >20-fold for a model with similar discrimination. Patient groups with such different outcome risks are unlikely to have similar benefit–harm tradeoffs for most therapies, even though they may be included in the same trial. EQRR = extreme-quartile risk ratio.

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

Most prediction models developed are not applied in clinical practice (81). Some fields produce many new prediction models without clear purpose except as publishable analytic exercises. The development of new models should instead start from some clinical need. Thoughtful selection of a "risk-sensitive" decision is a crucial step in developing a useful clinical prediction model (83). The use of a prediction model in clinical practice may be an implicit marker of 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$_2$ score (67, 68) (and its variations [84]), the atherosclerotic cardiovascular disease score (47), and chest pain tools (85, 86) 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 (such as cervical spine fracture [87], intracranial hemorrhage [88], and pulmonary embolism [89]) in low-risk patients to reduce the harms and burdens of further diagnostic testing is a marker of 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 reexamine the risk-specific benefits (or validate predictions of benefit) in new trial data are highly valuable.

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

The advantages of easily and reliably obtainable clinical characteristics as predictors should be obvious. Nevertheless, the literature contains many examples of models that include variables not ordinarily obtained in clinical care. For example, waist-to-hip ratio is a very strong predictor of diabetes and cardiovascular risk (90, 91), 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 questionable value for use in models (86). By raising the burden of variable ascertainment, we lower the probability that a prediction model will be used, compared with selecting the best treatment on average. Again, because of the abundance of published risk models, well-established risk predictors can usually be ascertained from the literature before the analysis of trial data, even when internal risk models will be used to stratify the data.

Explication of Excluded Criteria

The PATH technical expert panel did not reach consensus (as defined by a mean agreement score <3) on 2 additional criteria to identify when a risk-modeling approach is likely to be of value to analyze RCT results. The SDs of the agreement scores were also relatively high (>1) for these criteria, reflecting the conflicting positions held by panelists. Both criteria are described in the following paragraphs.

When the outcome rate is lower.

A low outcome rate is associated with a more asymmetrical 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 quartile of patients at lowest risk is anticipated to be approximately 1% (48). Although benefit is limited by a floor effect—it is impossible to decrease outcome risk to less than 0—treatment-related harms can still be substantial. The high-risk group (that is, the highest risk quartile) in these low-outcome trials frequently has outcome rates more than 10-fold those found in the low-risk group and may account for most of the benefit in the trial (**Figure 4**, *right*). These skewed distributions follow from the logistic regression scale (log odds) and Cox regression scale (log hazard [48]). This makes the average risk (and treatment benefit) misleading even for typical patients enrolled in the trial (48, 92). Nevertheless, the expert panelists disagreed about whether 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 2 treatments are clinically very different (for example, medicine vs. 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 symptomatic carotid stenosis (51), or for percutaneous coronary intervention versus coronary artery bypass grafting for non-acute coronary artery disease (93). Nevertheless, the technical expert panel disagreed on whether this criterion was a reliable marker of worthwhile opportunities for risk modeling. Indeed, when interventions in alternative trial groups differ substantially, individual variables may interact with treatment, making an effect-modeling approach more advantageous than a risk-modeling approach, such as in the SYNTAX (Synergy Between Percutaneous Coronary Intervention With Taxus and Cardiac Surgery) score II model for coronary artery bypass grafting versus percutaneous coronary intervention (93).

JUSTIFICATION OF GUIDANCE ON RISK-MODELING STRATEGIES TO IDENTIFY HTE

The following criteria describe best methodological practices in the conduct of risk-modeling approaches to identify HTE (**Figure 3** of the PATH Statement [26]). The motivations regarding these statements are elaborated, including reservations, considerations, and caveats.

Table 1. Mathematical Dependence of Treatment Effect on CER

Measure	Definition
Absolute risk difference	CER − TER
Relative risk reduction	1 − (TER/CER)
Odds ratio	[TER/(1 − TER)]/[CER/(1 − CER)]

CER = control event rate; TER = treatment event rate.

General

1. 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 an RCT. This rate may vary between 5- and even 20-fold across risk strata in trial populations (48, 94). 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 (nonlinearly related) contrasts of the same 2 quantities (control and treatment event rates), when the control event rate changes across subgroups, the treatment effect can remain constant on (at most) 1 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, 95). Thus, the widespread assumption that benefit–harm tradeoffs are usually similar for patients meeting trial enrollment criteria is demonstrably false and may be 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 underreporting of trial results (6).

2. 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 risk markers that may be important are considered. Thus, it is generally not possible to use 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 (that is, subgrouping) scheme (96, 97). In theory, these issues asymptotically diminish as databases become infinitely large, but clinical reasoning remains critical to the process of variable selection and model specification in the identification of a clinically plausible, clinically useful, and clinically usable model from the limited data sources that are 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

3. 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), risk-based analysis of HTE can often be done 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, definitions of predictor and outcome variables 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.

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 group only.

When high-quality, externally developed models are not available, internally derived (or endogenous) models can be used. Guidance on good prediction modeling practice should be followed—for example, many events per independent variable and prespecified, a priori selection of risk variables based on prior literature (98, 99). Models derived directly on RCT data may provide internally valid estimates of treatment effect within risk strata. One approach is to ignore treatment assignment in developing the model (100). 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 "testimation bias" [98]) that complicate effect modeling. Alternatively, some have recommended that only the control group be used to model risk (101) because this ostensibly generates the best estimate of the control event rate. However, modeling on the control group only can induce differential model fit on the 2 trial groups, biasing treatment effect estimates across risk strata and generally exaggerating HTE (49, 100, 102); various cross-validation techniques have been proposed to address this bias even for modeling on only the control group (103). Concerns about differential fit between groups from endogenously derived models may also apply when randomization is unbalanced (for example, 1:2 or 1:3 randomization) or when treatment effects are very large, such that the control group has many more events than the treatment group. When imbalance in events is caused by very large treatment effects, risk-based HTE may be less clinically relevant (Figure 3). Potential approaches to modeling on trials

Table 2. A Meta-research Agenda for Predictive Approaches to HTE

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 for the primary outcome vs. risk for treatment-related harm) 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, 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 dropout, loss to follow-up, poor adherence, and treatment switching.

Identify the appropriate clinical contexts for which modeling multiple dimensions of risks (e.g., risk for the primary outcome, risk for treatment-related harm, and risk for an important competing outcome) is important and feasible for adequate disaggregation of benefit–harm tradeoffs.

Determine methods to analyze trials with multiple important outcomes, or outcomes where differences in treatment effect may be related to choice of follow-up time.

Better understand the effect 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 groups 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 best to extend these approaches to other trial designs (e.g., longitudinal studies and dynamic treatment regimens).

HTE = heterogeneity of treatment effects.

with imbalanced randomization or strong treatment effects should be evaluated in future research (Table 2).

5. When developing new risk models or updating externally developed risk models, specify the analytic data plan before examining trial data and follow guidance for best practices for prediction model development.

Although not the focus of the PATH Statement, the development of new risk models or updating of externally developed models is supported by existing guidance (98, 99, 104). In particular, the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) explanation and elaboration document (104) offers detailed, referenced guidance (with published exemplars) on how to design, conduct, and analyze prediction model studies with a view to limiting risk of bias and maximizing the clinical usefulness of the model. 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 showing clinical trial data to readers in subgroups, (which is part of the guidance under Apply the Model, and Report Results). The TRIPOD explanation and elaboration document (104) 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 excluding patients who were randomly assigned to a group. Alternative approaches for subgroup identification in the presence of missing variables should be investigated in future research (Table 2). The PROGRESS (Prognosis Research Strategy) series of articles (105–108) and several textbooks also offer guidance on the optimal development of clinical prediction models (98, 99).

Because 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 practices for individual patient meta-analysis (109). In particular, inclusion of study-specific intercepts to account for unexplained risk heterogeneity is recommended (110, 111). 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 (112, 113).

Apply the Model, and Report Results

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

7. Report distribution of predicted risk (or the risk score) in each group of the trial and in the overall study population.

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

9. When there are important treatment-related harms, these harms should be reported in each risk stratum to support stratum-specific evaluation of benefit–harm tradeoffs.

Consistent with the TRIPOD Statement, we recommend that measures of discrimination and calibration be presented whether an externally or internally derived model is applied (recommendation 6 in **Figure 3** of the PATH Statement [26]). However, these conventional measures of model performance should not be confused 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 (114), CHA_2DS_2-VASc (84), and ABCD2 (115), may be useful for trial risk stratification but do not yield predictions for calibration.

Although its importance was highlighted 2 decades ago (94), reporting the distribution of baseline risk is rarely done (recommendation 7 in **Figure 3** of the PATH Statement [26]). Thus, assessing the degree of baseline risk heterogeneity is generally impossible in most published clinical trials. Risk reporting should allow readers to assess the full distribution of risk in the study population, either graphically or through information on the mean, SD, 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 attributes for participants in each study group) should also include the population mean and median predicted baseline risk (or risk score) with measures of variability, as well as additional information on the population distribution of risk if skewness is substantial (such as quartiles or percentiles, a histogram, or a box plot). If the study population is largely homogeneous with regard to overall risk, the reader will know that generalizing the study results to populations with substantially different risk would be speculative. If heterogeneity is substantial in the study population, reviewers will know that conducting a risk-stratified analysis is particularly important.

As an initial step, we recommend grouping patients using quantiles (such as quartiles) for reporting purposes and showing and estimating treatment effects separately in these groups (for example, dividing patients into equal-sized quarters) (recommendation 8 in **Figure 3** of the PATH Statement [26]). Reporting treatment effects across strata is important because it illustrates how the absolute risk difference varies across the study population, regardless of whether the *relative* effect is constant (**Figure 4** [*right*] is an example). In addition, it permits evaluation of the assumption of a constant relative effect across risk strata (see recommendation 10 in **Figure 3** of the PATH Statement [26] and explanation in discussion under recommendation 10 here). Alternatively, treatment effects can be presented by continuous risk, as seen in **Figure 2**, rather than by quantiles (which are sample-dependent). As discussed earlier, examining variation in relative treatment effects may be particularly important when even a small amount of treatment-related harm exists (3, 66). In 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 such horizons). Absolute treatment effect estimates can be summarized by cumulative incidences at a clinically meaningful time point (or several such points). In reporting risk-stratified results, authors provide readers with the information needed to easily determine the amount of variation in risk difference or number needed to treat 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 (recommendation 3 in **Figure 4** of the PATH Statement [26]).

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 these 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, subgroups stratified by risk for the primary outcome cannot be assumed to have similar rates of treatment-related harms. For example, patients with atrial fibrillation who have higher $CHADS_2$ scores (indicating higher stroke risk and greater potential benefit from anticoagulation) also have substantially higher risk for bleeding (72). Patients with higher risk for stroke recurrence, according to a recurrence risk score, may benefit more from pioglitazone but also have higher risk for pioglitazone-related bone fracture (65). Because of the potential correlation between these 2 risk dimensions (that is, between risk for the primary outcome and risk for treatment-related harm), event rates for these harms should be presented at a level of disaggregation that is congruent with that of the primary outcome so that readers can determine benefit–harm tradeoffs within risk strata (recommendation 9 in **Figure 3** of the PATH Statement [26]).

Risk modeling may be most useful when predictors of risk for the primary outcome and benefits of therapy are poorly (or negatively) correlated with risk for treatment-related harm. This will maximize heterogeneity in the benefit–harm tradeoffs across risk strata, increasing the decisional value of the risk model. Although several investigators have sought to arithmetically combine separate prediction models for outcome risk and treatment-related harm to stratify trial results by benefit–harm tradeoff (51, 69, 71), this approach can be very sensitive to miscalibration of the 2 models (which may compound miscalibration of benefit–harm tradeoffs). The best approach to modeling benefits and harms simultaneously is beyond the scope of these recommendations and is an important topic for future research (**Table 2**).

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

Although testing for a statistical interaction between subgroups is recommended to determine HTE, when the outcome rates vary substantially across strata, the risk difference can also be assumed to vary (**Table 1** and 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 because a nonsignificant result is far more likely to reflect low power than true consistency of effects on the risk difference scale. Statistically testing a risk-by-treatment interaction on the relative scale (for example, 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. Nevertheless, 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 a risk-by-treatment interaction test, using a continuous measure of risk (such as the logit of risk) typically provides superior power compared with testing for effect differences across distinct risk groups (Figure 4, *middle right*) (116). A visual (nonparametric) exploration of how the relative effect varies across values of outcome risk may ensure the appropriateness of linear effect modification. Testing for a nonlinear interaction between risk and treatment (for example, using the logit of risk in a quadratic term, or with another flexible nonlinear shape [117, 118]) may also be useful. However, 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 the existence of an overall treatment effect is established, determining the risk-specific treatment effect should be considered an estimation problem (rather than a hypothesis-testing problem). Flexibly modeling the treatment effect across risk strata, or simply reporting the effects across subgroups defined by quantiles (such as quartiles), provides useful information regardless of the *P* value of the interaction terms testing effect modification on the relative scale. Standard errors across levels of risk can be estimated through a proportional interactions model (119, 120). Most important, the presence or absence of a statistically significant treatment interaction term (on the relative scale) should not be conflated with the presence or absence of clinically important HTE (on the absolute scale) (see discussion of recommendation 1 under Justification of Caveats and Considerations Before Moving to Clinical Practice, below).

Table 3. Hypothetical Example Presentation of the Effects of Model-Based Decision Making

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

JUSTIFICATION OF CAVEATS AND CONSIDERATIONS BEFORE MOVING TO CLINICAL PRACTICE

The following considerations relate to the translation of findings from predictive approaches to HTE analyses into clinical practice (Figure 4 in the PATH Statement [26]). Clinical translation of these analyses is a complex topic that includes 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, these analytic considerations are priorities 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.

The clinical significance of HTE should generally be discussed with reference to the absolute scale, whereas a relative scale (such as 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 (for example, number of patients treated or number of events avoided with vs. without model use) and should consider decision analytic approaches for evaluation (58, 121). Table 3 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-stratum 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 (107). 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. Of note, external validity is a general concern for RCT results and their subgroup analyses and is not confined to results subgrouped using prediction models (122).

3. Clinical implementation may be supported by translating multivariable risk-based subgroup analysis into models yielding continuous treatment effect predictions to avoid artifactual 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 (123). Categorization into risk groups suggests that risk and treatment effects are homogeneous within groups and leads to a potentially misleading "step function" in estimation of risk or treatment effect. 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 anticipated benefit. In addition, quantiles have specific disadvantages in that they are sample-driven cut points, which leads to difficulties in comparing results across studies (124) and may also obscure problems with model calibration. For example, use of an internally developed risk model and use of the Framingham model to stratify patients in the DPP trial seemed to yield near-identical results (32, 33). However, if the trial population had been divided into groups based on predicted risk thresholds (as it would be in clinical practice), risk groups defined by the Framingham model would have shown that the Framingham model was poorly calibrated to the DPP trial population. Thus, although 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 (**Figure 1**, *legend*).

TREATMENT EFFECT MODELING TO IDENTIFY HTE

Considerations Regarding the Inclusion of Rigorously Selected Effect Modifiers

Conventional 1-variable-at-a-time subgroup analyses 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 (**Figure 5** in the PATH Statement [26]).

Although relative effect modifiers are difficult to reliably identify, they are highly influential on individual patient predictions of benefit (122). Including spurious false-positive interaction terms in models that predict treatment benefit can mistarget therapies, and excluding true interaction terms limits the usefulness of prediction by substantially lowering discrimination of patients who benefit from those who do not (30). Whether to include a treatment effect interaction term in a prediction model is a fraught and consequential decision; the PATH technical expert panel accordingly recommends a cautious approach. We restrict our recommendations to the unusual situation where highly credible effect modifiers have been identified, and we 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 and should not rely solely on statistical criteria (such as P *value thresholds).*

Important efforts have been made to establish criteria that might identify highly credible subgroup analyses (9, 19, 125). A newly proposed tool (ICEMAN [Instrument for assessing the Credibility of Effect Modification ANalyses] [125]) is based on 5 criteria to evaluate credibility of effect modification. Four of these criteria are related to prespecification and markers of the "prior probability" or plausibility of effect modification: presence of prior evidence, prespecification of a few primary subgroup analyses, prespecification of the anticipated directionality of effect modification, and 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 for acute myocardial infarction or acute ischemic stroke (126, 127), gender as a modifier of the effect of thiazolidinediones on fracture risk (128–130), and urinary protein excretion as a modifier of the effect of angiotensin-converting enzyme inhibition on the progression of chronic kidney disease (77, 131).

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

Prespecification of primary subgroups should include explicit definitions and categories of the subgroup variables, including cutoff thresholds for continuous or ordinal variables where these are used and the anticipated direction of effect modification. When primary subgroup analyses are few in number, fully prespecified, hypothesis-driven, and statistically robust (that is, 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 to generate hypotheses, which can then be tested (and usually disproved [23, 25, 132, 133]) in future studies.

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 tri-

Table 4. Methodological Literature on the Conduct of Regression-Modeling Approaches to HTE Analysis

Approach	Description
Risk-modeling approaches (7, 48, 66, 100, 119, 120, 122, 151–154)	Using a multivariable risk model developed blinded to treatment effect, analyze the relationship between baseline risk and treatment effect on the relative and absolute scales. Whereas 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 (122, 147–150, 154–161) Subgroup identification (2-step process) (147, 148, 155) Individualized treatment effects (1-step process) (122, 149, 154, 156–158)	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 (136–146)	Classify patients into those who benefit from treatment (positive individualized treatment effect) and those who do not (negative individualized treatment effect) through identification of modifiers of treatment effects on the relative scale.

HTE = heterogeneity of treatment effects.

als of conventional size (even when these are pooled), treatment interaction terms selected for inclusion are likely to overestimate the true interaction effects (that is, from overfitting) (4, 30, 134). Therefore, even when only highly credible interaction terms are included, we recommend model-building procedures that take into account model complexity (that is, approaches using regularization or penalization) whenever interactions are included (recommendation 3 in **Figure 5** of the PATH Statement [26]).

Caveats and Considerations for Data-Driven Effect Modeling

Emerging "data-driven" methods—proposed to develop effect models on trial data when prior information on effect modifiers is limited—are a promising area of research (28), but the technical expert panel believed them to be at too formative a stage to offer recommendations. A systematic scoping review (135) was done in an effort to characterize this rapidly evolving literature, and it revealed future research opportunities. **Table 4** 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—that is, based only on the sign of the effect. Because these are classification rather than prediction methods, we do not discuss them in our guidance but direct interested readers to the literature on these approaches (136–146). 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 (98, 147, 148). Alternatively, various methods of penalization or regularization to reduce the likelihood of overfitting have been proposed and tested (149). Although the PATH technical expert panel did not articulate a full set of methodological best practices for treatment effect modeling given more limited practical experience (compared with risk modeling), we offer caveats and considerations for this type of predictive HTE analysis (and their justifications) in the following paragraphs.

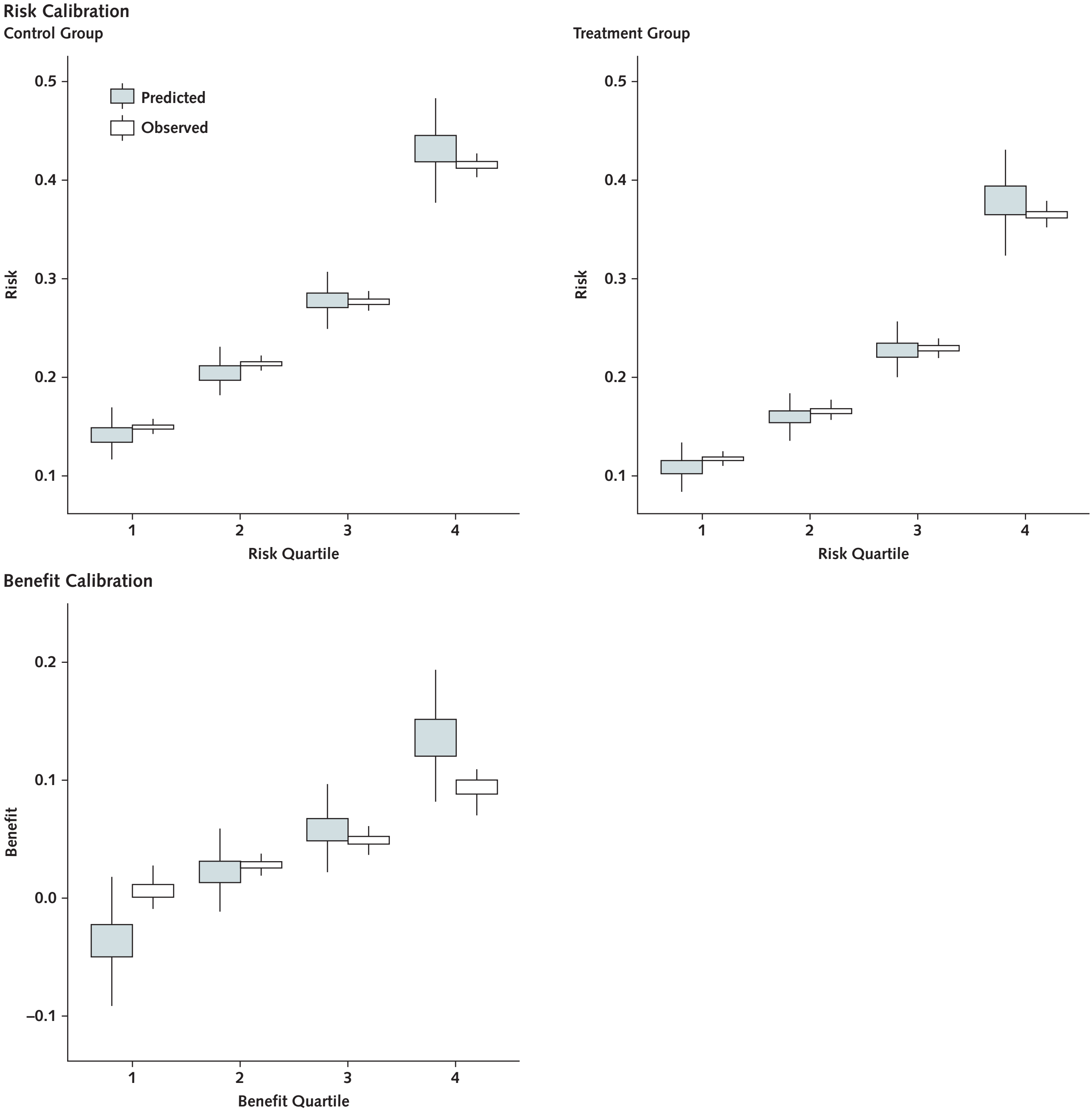
2. Avoid 1-variable-at-a-time null hypothesis testing or stepwise selection (such as backward selection or forward selection) strategies to select single-variable relative effect modifiers.

One-variable-at-a-time null hypothesis testing will preferentially select effects that are overestimated within the sample database (that is, 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 (recommendation 1 in **Figure 5** of the PATH Statement [26]). When multiple relative effect modifiers are hypothesized to be of potential importance in determining treatment effects, the value of including these interactions can be assessed simultaneously by a single overall test, limiting the opportunity for type I error and testimation bias (104). To increase the power of this test, the number of treatment effect interactions included should be limited (104). 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 (that is, including interaction terms) will not add substantially to a risk-modeling approach.

3. Avoid the use of regression methods that do not take into account model complexity when estimating coefficients (for example, "conventional" unpenalized maximum-likelihood regression) when 1 or more treatment-by-covariate interaction terms are included in a treatment effect model.

Although no consensus exists on the optimal approach for including relative effect modifiers in a prediction model (that is, an "effect model"), conventional regression techniques will generally result in overfitting. Penalized estimation (such as with LASSO [least absolute shrinkage and selection operator] regression or ridge regression, elastic net regularization regression, Bayesian penalization, and other non–regression-based methods discussed earlier) at least partially addresses the tendency to overfit benefit predictions. The optimal ap-

Figure 6. Evaluating model performance: a comparison of conventional outcome risk calibration in control and treatment groups vs. benefit calibration.



These data are box plots of predicted and observed hypothetical examples of event rates divided by quartiles of predicted risk in the control and treatment groups of a hypothetical randomized controlled trial (500 simulations) (*top*). These rates seem to demonstrate appropriate model calibration. However, examining the same data for predicted and observed benefit (differences in event rates) by quarters of predicted benefit (*bottom*) reveals very poor model calibration at the extreme quarters. This poor calibration occurs because miscalibration for the risk difference includes error from both control and treatment groups and because the scale of risk difference is much smaller than that of outcome risk. These data were 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 the Tukey 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.

proach 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 data sets (or subsets) for variable and model selection for determining the subgrouping (that is, reference class) scheme and for treatment effect estimation can also avoid or mitigate testimation bias.

4. Avoid evaluating models that predict treatment benefit using only conventional metrics for outcome prediction (for example, 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, not for their ability to predict outcome risk. Calibration for outcome risk can be seriously misleading in evaluation of models that purport to predict treatment benefit (**Figure 6**). The discrepancy arises because benefit miscalibration compounds the error in risk estimation in the control and treatment groups and magnifies this error (that is, the scale of risk difference is typically much smaller than that of outcome rate). Evaluation methods that pertain to models intended for treatment selection or benefit prediction are discussed further in the following section.

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 (such as the c-statistic) does not 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 because only 1 of the possible outcomes is observed for each patient (the actual outcome they experienced under the treatment to which they were randomly assigned and not the counterfactual outcome under the alternative) (162).

For example, for "predictive" biomarkers (factors that can aid in treatment selection), some statisticians have suggested evaluating performance by the sensitivity and specificity of the biomarker for benefit rather than risk (163). However, these quantities can be estimated 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 (benefit [bad outcome without treatment and good outcome with treatment {1,0}], neutral [good or bad outcome regardless of treatment {0,0 or 1,1}], or harm [good outcome without treatment and bad outcome with treatment {0,1}]) (164). For example, 1 proposed method assumes that no participants are harmed by treatment (165); others assume that, conditional on a set of covariates, the potential outcomes with and without treatment are independent (150, 166). Without such assumptions, we can focus only on a model's ability to predict outcome risk in 1 group of a trial or the other, rather than the model's ability to predict benefit (the difference in outcome across groups) (164, 167).

Because 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. The c-statistic, commonly used to measure discrimination in outcome risk models, has recently been adapted to evaluate treatment effect prediction (150). To do this evaluation, 2 patients discordant on treatment assignment are matched according to their predicted benefit (that is, 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 following 3 benefit categories according to their "observed benefit" based on a comparison of outcomes in the control and treated patients: benefit (1,0), neutral (1,1 or 0,0), or harm (0,1). The c-statistic assesses how well the model discriminates pairs of patients on the basis of this trinary "outcome." Again, the definition of "observed benefit" assumes that the potential outcomes with the 2 therapies are independent within each patient. Because the potential outcomes within each patient are presumably dependent to some (unknowable) degree, the "observed benefit" contains more randomness than the actual (unobservable) treatment benefit for each individual patient. This leads to conservative estimates of the c-statistic.

The usefulness of a model depends on its ability not only to accurately predict within-strata treatment effect but also to improve decisions. Of course, the ultimate test of a predictive approach is to compare decisions (or outcomes) in settings that use such individualized predictions to guide care in an experiment (168). 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 effect of models on clinical decision making. Evaluations of clinical usefulness depend on model performance relative to a specific decision threshold—that is, the absolute risk difference that perfectly balances the burdens, harms, and costs of therapy. Decision curve analysis (59) has been proposed to evaluate the clinical usefulness of prediction models in 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 mathematically derive a utility weight of benefits versus harms (implicitly revealed by the selection of that threshold). The approach has also been adapted to evaluate the potential effect of prediction of treatment benefit on decision making compared with the default best overall strategy (that is, treat all or treat none) (121).

Because identifying the correct treatment for any given individual is impossible (except, with assumptions, in *n*-of-1 trials [27, 169]), all of these methods evaluate evidence personalization *indirectly* by evaluating whether a particular prediction–decision strategy optimizes benefits for a population (58)—which occurs 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 most large, phase 3 clinical trials (170–173). 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 multiperson *n*-of-1—trials, which some consider the only means of estimating "person-level" treatment effects. We anticipate that observational studies will play an increasingly important role in studying both treatment effects and HTE, but 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 [**Table 2**]). Although each of the approaches we describe is consistent with the broad goal of evidence personalization, the methods are sufficiently distinct to be beyond the scope of this statement. Notwithstanding the limitations, we emphasize that the PATH Statement applies to the comparison of treatments as well as the comparison of treatment versus 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 best to cope with postrandomization events, including dropout, nonadherence, 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 expert panel report on prevention and treatment of missing data in clinical trials (174, 175) highlighted the need to clearly define objectives and estimands, and since the subsequent ICH E9(R1) draft addendum (176). We direct readers to recent literature on this topic (177–179). In general, the primary analysis of most trials is often an intention-to-treat analysis. However, other contrasts are also of clinical import and interest. In particular, the direct causal effect of treatment (that is, the effect *if* a patient adheres to treatment, estimated with a per protocol or adherence-adjusted analysis [180–184]) 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 be done only with methods based on unverifiable assumptions; misspecifying a model predicting nonadherence (or using an instrumental variable approach when causes of nonadherence or dropout are complex) can lead to biased estimates of treatment effects. An intention-to-treat analysis is generally believed to yield an unbiased estimate of the treatment policy, although this may be less appropriate for shared decision making. More research is needed regarding optimal ways to combine predictive HTE approaches with approaches that estimate direct treatment or adherence-adjusted effects.

DISCUSSION

The PATH Statement comprises 4 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, 185). Developers of the PATH Statement recognize the inherent difficulties and fundamental limitations of using group data to estimate treatment effects in individuals and have enumerated some of these challenges (26). As more deeply explored in this explanation and elaboration document, there remain substantial barriers to a full understanding of the potential of predictive HTE approaches (186). **Table 2** 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 ensure that incorporating these methods does not cause more harm than benefit. We also need research to better integrate clinical prediction into practice (187), to understand how to individualize clinical practice guidelines, to establish or extend reporting guidelines (188), to establish new models of data ownership to facilitate individual-level meta-analyses (189), and to reengineer the clinical research infrastructure to support substantially larger, clinically integrated trials that are sufficiently powered to determine HTE (or to develop our ability to predict when observational data are likely to be sufficiently debiased for reliable HTE determination) (190). Many recent and ongoing organizational and technical advances should help toward this evolution (189, 191–193).

Because the PATH Statement focused on prediction in randomized trials, we did not explore the use of observational data—and when they may be sufficiently debiased for reliable identification of HTE (10, 194, 195). In addition, there is an evolving set of tools for data-driven approaches to predicting patient benefit, including machine-learning techniques (28, 196, 197). The PATH Statement should thus be understood as a formative first step (along a much longer path) toward the goal of personalized predictions of treatment benefit using the best available evidence.

From Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts (D.M.K., J.K.P., J.B.W.); Erasmus Medical Center, Rotterdam, the Netherlands, and Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts (D.V.); Boston University, Boston, Massachusetts (R.D.); Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (S.G., J.P.I.); University of Michigan, Ann Arbor, Michigan (R.H.); Duke Clinical Research Institute, Duke University, Durham, North Carolina (B.P., M.P.); Virginia Polytechnic Institute and State University, Blacksburg, Virginia (S.M.); Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, Massachusetts (G.R.); Schools of Medicine and Public Health, Yale University, New Haven, Connecticut (J.S.R.); Center for Cardiovascular Health Services Research, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, and Tufts Clinical and Translational Science Institute, Boston, Massachusetts (H.P.S.); Center on Aging and Health, Johns Hopkins University, Baltimore, Maryland (R.V.); Memorial Sloan Kettering Cancer Center, New York, New York (A.V.); Tufts Medical Center, Boston, Massachusetts; and Leiden University Medical Center, Leiden, the Netherlands (E.W.S.).

Disclaimer: The views, statements, and opinions presented in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors, or its Methodology Committee.

Acknowledgment: The authors thank Mark Adkins, Teddy Balan, and Dan Sjoberg for excellent technical support in analyses included in figures and the appendix tables that support the PATH Statement (26). They also thank the *Annals of Internal Medicine* editors and reviewers, whose thoughtful feedback greatly improved this work. They thank Jennifer Lutz and Christine Lundquist for assistance with copyediting and creating exhibits.

Financial Support: Development of the PATH Statement was supported through contract SA.Tufts.PARC.OSCO.2018.01.25 from the PCORI Predictive Analytics Resource Center. This work was also informed by a 2018 conference ("Evidence and the Individual Patient: Understanding Heterogeneous Treatment Effects for Patient-Centered Care") convened by the National Academy of Medicine and funded through a PCORI Eugene Washington Engagement Award (1900-TMC).

Disclosures: Dr. Kent reports grants from PCORI during the conduct of the study. Dr. Goodman reports personal fees from PCORI outside the submitted work. Dr. Pencina reports grants from PCORI (Tufts Subaward) during the conduct of the study; grants from Sanofi/Regeneron, Amgen, and Bristol-Myers Squibb outside the submitted work; and personal fees from Boehringer Ingelheim and Merck outside the submitted work. Dr. Ross reports personal fees from PCORI during the conduct of the study and grants from the U.S. Food and Drug Administration, Medtronic, Johnson & Johnson, the Centers for Medicare & Medicaid Services, Blue Cross Blue Shield Association, the Agency for Healthcare Research and Quality, the National Institutes of Health (National Heart, Lung, and Blood Institute), and Laura and John Arnold Foundation outside the submitted work. Dr. Varadhan reports personal fees from Tufts University during the conduct of the study. Dr. Vickers reports grants from the National Institutes of Health during the conduct of the study. Dr. Wong reports grants from PCORI during the conduct of the study. Dr. Steyerberg reports royalties from Springer for his book *Clinical Prediction Models*. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M18-3668.

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

Current Author Addresses: Drs. Kent, Paulus, Raman, and Selker: Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, 800 Washington Street, Box 63, Boston, MA 02111.

Dr. van Klaveren: Erasmus University Medical Center, Doctor Molewaterplein 40, 3015 GD Rotterdam, the Netherlands.

Dr. D'Agostino: Boston University Mathematics and Statistics Department, 111 Cummington Street, Boston, MA 02215.

Dr. Goodman: Stanford University School of Medicine, 150 Governor's Lane, Room T265, Stanford, CA 94305.

Dr. Hayward: VA Ann Arbor Health Services Research and Development, 2800 Plymouth Road, Building 14, G100-36, Ann Arbor, MI 48109.

Dr. Ioannidis: Stanford Prevention Research Center, 1265 Welch Road, Stanford, CA 94305.

Ms. Patrick-Lake: Evidation Health, 167 2nd Avenue, San Mateo, CA 94401.

Dr. Morton: Virginia Tech, North End Center Suite 4300, 300 Turner Street NW, Blacksburg, VA 24061.

Dr. Pencina: Duke Clinical Research Institute, 200 Trent Street, Durham, NC 27710.

Dr. Ross: Yale University School of Medicine, PO Box 208093, New Haven, CT 06520.

Dr. Varadhan: Johns Hopkins University, Division of Biostatistics and Bioinformatics, 550 North Broadway, Suite 1103-A, Baltimore, MD 21205.

Dr. Vickers: Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, NY 10017.

Dr. Wong: Tufts Medical Center, 800 Washington Street #302, Boston, MA 02111.

Dr. Steyerberg: Erasmus University Medical Center, PO Box 2040, 3055 PC Rotterdam, the Netherlands.

Author Contributions: Conception and design: D.M. Kent, J.K. Paulus, R. Hayward, J.P.A. Ioannidis, B. Patrick-Lake, J.S. Ross, A. Vickers, J.B. Wong, E.W. Steyerberg.

Analysis and interpretation of the data: D.M. Kent, J.K. Paulus, R. D'Agostino, R. Hayward, J.P.A. Ioannidis, R. Varadhan, J.B. Wong, E.W. Steyerberg.

Drafting of the article: D.M. Kent, J.K. Paulus, R. D'Agostino, S. Goodman, J.P.A. Ioannidis, A. Vickers, J.B. Wong.

Critical revision of the article for important intellectual content: D.M. Kent, D. van Klaveren, J.K. Paulus, R. D'Agostino, S. Goodman, R. Hayward, J.P.A. Ioannidis, S. Morton, M. Pencina, G. Raman, J.S. Ross, H.P. Selker, R. Varadhan, A. Vickers, J.B. Wong, E.W. Steyerberg.

Final approval of the article: D.M. Kent, D. van Klaveren, J.K. Paulus, R. D'Agostino, S. Goodman, R. Hayward, J.P.A. Ioannidis, B. Patrick-Lake, S. Morton, M. Pencina, G. Raman, J.S. Ross, H.P. Selker, R. Varadhan, A. Vickers, J.B. Wong, E.W. Steyerberg.

Provision of study materials or patients: D.M. Kent, J.B. Wong.

Statistical expertise: D.M. Kent, D. van Klaveren, R. D'Agostino, R. Hayward, J.P.A. Ioannidis, S. Morton, R. Varadhan, A. Vickers, J.B. Wong, E.W. Steyerberg.

Obtaining of funding: D.M. Kent, J.K. Paulus, J.B. Wong.

Administrative, technical, or logistic support: D.M. Kent, J.K. Paulus, G. Raman, H.P. Selker, J.B. Wong.

Collection and assembly of data: D.M. Kent, J.K. Paulus, G. Raman, J.B. Wong.

References

1. **Rothwell PM.** Can overall results of clinical trials be applied to all patients? Lancet. 1995;345:1616-9. [PMID: 7783541]
2. **Rothwell PM, Mehta Z, Howard SC, et al.** Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. Lancet. 2005;365:256-65. [PMID: 15652609]
3. **Kent DM, Hayward RA.** Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007;298:1209-12. [PMID: 17848656]
4. **Kent DM, Steyerberg E, van Klaveren D.** Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;363:k4245. [PMID: 30530757] doi:10.1136/bmj.k4245
5. **Kravitz RL, Duan N, Braslow J.** Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82:661-87. [PMID: 15595946]
6. **Hayward RA, Kent DM, Vijan S, et al.** Reporting clinical trial results to inform providers, payers, and consumers. Health Aff (Millwood). 2005;24:1571-81. [PMID: 16284031]
7. **Kent DM, Rothwell PM, Ioannidis JP, et al.** Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11:85. [PMID: 20704705] doi:10.1186/1745-6215-11-85
8. **Varadhan R, Segal JB, Boyd CM, et al.** A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2013;66:818-25. [PMID: 23651763] doi:10.1016/j.jclinepi.2013.02.009
9. **Sun X, Ioannidis JP, Agoritsas T, et al.** How to use a subgroup analysis: users' guide to the medical literature. JAMA. 2014;311:405-11. [PMID: 24449319] doi:10.1001/jama.2013.285063
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;45:2184-93. [PMID: 27864403] doi:10.1093/ije/dyw125
11. **Davidoff F.** Can knowledge about heterogeneity in treatment effects help us choose wisely? Ann Intern Med. 2017;166:141-2. [PMID: 27820948] doi:10.7326/M16-1721
12. **Lagakos SW.** The challenge of subgroup analyses–reporting without distorting. N Engl J Med. 2006;354:1667-9. [PMID: 16625007]
13. **Rothwell PM.** Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176-86. [PMID: 15639301]
14. **Hernández AV, Boersma E, Murray GD, et al.** Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J. 2006;151:257-64. [PMID: 16442886]
15. **Wang R, Lagakos SW, Ware JH, et al.** Statistics in medicine–reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357:2189-94. [PMID: 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:I98-101. [PMID: 6133654]
17. **Tannock IF.** False-positive results in clinical trials: multiple significance tests and the problem of unreported comparisons. J Natl Cancer Inst. 1996;88:206-7. [PMID: 8632495]
18. **Assmann SF, Pocock SJ, Enos LE, et al.** Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000;355:1064-9. [PMID: 10744093]
19. **Oxman AD, Guyatt GH.** A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116:78-84. [PMID: 1530753]
20. **Pocock SJ, Assmann SE, Enos LE, et al.** Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med. 2002;21:2917-30. [PMID: 12325108]
21. **Stallones RA.** The use and abuse of subgroup analysis in epidemiological research. Prev Med. 1987;16:183-94. [PMID: 3295858]
22. **Parker AB, Naylor CD.** Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J. 2000;139:952-61. [PMID: 10827374]
23. **Brookes ST, Whitley E, Peters TJ, et al.** Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess. 2001;5:1-56. [PMID: 11701102]
24. **Brookes ST, Whitely E, Egger M, et al.** Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57:229-36. [PMID: 15066682]
25. **Burke JF, Sussman JB, Kent DM, et al.** Three simple rules to ensure reasonably credible subgroup analyses. BMJ. 2015;351:h5651. [PMID: 26537915] doi:10.1136/bmj.h5651
26. **Kent DM, Paulus JK, van Klaveren D, et al.** The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. Ann Intern Med. 12 November 2019. [Epub ahead of print]. doi:10.7326/M18-3667
27. **Raman G, Balk EM, Lai L, et al.** Evaluation of person-level heterogeneity of treatment effects in published multiperson N-of-1 studies: systematic review and reanalysis. BMJ Open. 2018;8:e017641. [PMID: 29804057] doi:10.1136/bmjopen-2017-017641
28. **Lipkovich I, Dmitrienko A, B R D'Agostino Sr.** Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med. 2017;36:136-96. [PMID: 27488683] doi:10.1002/sim.7064
29. **VanderWeele TJ, Luedtke AR, van der Laan MJ, et al.** Selecting optimal subgroups for treatment using many covariates. Epidemiology. 2019;30:334-41. [PMID: 30789432] doi:10.1097/EDE.0000000000000991
30. **van Klaveren D, Balan TA, Steyerberg EW, et al.** Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol. 2019;114:72-83. [PMID: 31195109] doi:10.1016/j.jclinepi.2019.05.029
31. **Greenland S, Rothman KJ, Lash TL.** Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
32. **Sussman JB, Kent DM, Nelson JP, et al.** Improving diabetes prevention with benefit based tailored treatment: risk based reanalysis of Diabetes Prevention Program. BMJ. 2015;350:h454. [PMID: 25697494] doi:10.1136/bmj.h454
33. **Wilson PW, Meigs JB, Sullivan L, et al.** Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007;167:1068-74. [PMID: 17533210]
34. **VanderWeele TJ, Robins JM.** The identification of synergism in the sufficient-component-cause framework. Epidemiology. 2007;18:329-39. [PMID: 17435441]
35. **VanderWeele TJ, Knol MJ.** A tutorial on interaction. Epidemiol Methods. 2014;3:33-72. doi:10.1515/em-2013-0005

36. **Hallqvist J, Ahlbom A, Diderichsen F, et al.** How to evaluate interaction between causes: a review of practices in cardiovascular epidemiology. J Intern Med. 1996;239:377-82. [PMID: 8642229]

37. **Andersson T, Alfredsson L, Källberg H, et al.** Calculating measures of biological interaction. Eur J Epidemiol. 2005;20:575-9. [PMID: 16119429]

38. **Ahlbom A, Alfredsson L.** Interaction: a word with two meanings creates confusion [Editorial]. Eur J Epidemiol. 2005;20:563-4. [PMID: 16119427]

39. **VanderWeele TJ, Robins JM.** Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology. 2007; 18:561-8. [PMID: 17700242]

40. **VanderWeele TJ, Robins JM.** Empirical and counterfactual conditions for sufficient cause interactions. Biometrika. 2008;95:49-61. doi:10.1093/biomet/asm090

41. **VanderWeele TJ, Knol MJ.** Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. Ann Intern Med. 2011;154:680-3. [PMID: 21576536] doi:10.7326/0003-4819-154-10-201105170-00008

42. **Harrell F, Lazzeroni L.** EHRs and RCTs: outcome prediction vs. optimal treatment selection. 2017. Accessed at www.fharrell.com/post/ehrs-rcts on 1 May 2019.

43. **Harrell F.** Viewpoints on heterogeneity of treatment effect and precision medicine. 2018. Accessed at www.fharrell.com/post/hteview on 1 May 2019.

44. **Engels EA, Schmid CH, Terrin N, et al.** Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19:1707-28. [PMID: 10861773]

45. **Sterne JA, Egger M.** Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol. 2001;54: 1046-55. [PMID: 11576817]

46. **Deeks JJ, Altman DG.** Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG, eds. Systematic Reviews in Health Care: Meta-Analysis in Context. 3rd ed. London: BMJ Publishing Group; 2003.

47. **Goff DC Jr, Lloyd-Jones DM, Bennett G, 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;63:2935-59. [PMID: 24239921] doi:10.1016/j.jacc.2013.11.005

48. **Kent DM, Nelson J, Dahabreh IJ, et al.** Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol. 2016;45:2075-88. [PMID: 27375287] doi:10.1093/ije/dyw118

49. **van Klaveren D, Balan TA, Steyerberg EW, et al.** Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. J Clin Epidemiol. 2019;114: 72-83. [PMID: 31195109] doi:10.1016/j.jclinepi.2019.05.029

50. **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;34:464-7. [PMID: 12574561]

51. **Rothwell PM, Warlow CP; European Carotid Surgery Trialists' Collaborative Group.** Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. Lancet. 1999;353: 2105-10. [PMID: 10382694]

52. **Frommeyer G, Eckardt L.** Drug-induced proarrhythmia: risk factors and electrophysiological mechanisms. Nat Rev Cardiol. 2016; 13:36-47. [PMID: 26194552] doi:10.1038/nrcardio.2015.110

53. **Roden DM.** Mechanisms and management of proarrhythmia. Am J Cardiol. 1998;82:49I-57I. [PMID: 9737654]

54. **Costa F, van Klaveren D, James S, et al; PRECISE-DAPT Study Investigators.** 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;389:1025-34. [PMID: 28290994] doi:10.1016/S0140-6736(17)30397-5

55. **Viscoli CM, Kent DM, Conwit R, et al; IRIS Trial Investigators.** Scoring system to optimize pioglitazone therapy after stroke based on fracture risk. Stroke. 2018:STROKEAHA118022745. [PMID: 30580725] doi:10.1161/STROKEAHA.118.022745

56. **Baker SG, Cook NR, Vickers A, et al.** Using relative utility curves to evaluate risk prediction. J R Stat Soc Ser A Stat Soc. 2009;172:729-48. [PMID: 20069131]

57. **Pauker SG, Kassirer JP.** The threshold approach to clinical decision making. N Engl J Med. 1980;302:1109-17. [PMID: 7366635]

58. **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. [PMID: 26810254] doi:10.1136/bmj.i6

59. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565-74. [PMID: 17099194]

60. **Baker SG, Kramer BS.** Evaluating a new marker for risk prediction: decision analysis to the rescue. Discov Med. 2012;14:181-8. [PMID: 23021372]

61. **Hammadah M, Kim JH, Tahhan AS, et al.** Use of high-sensitivity cardiac troponin for the exclusion of inducible myocardial ischemia: a cohort study. Ann Intern Med. 2018;169:751-60. [PMID: 30398528] doi:10.7326/M18-0670

62. **Baker SG.** Putting risk prediction in perspective: relative utility curves. J Natl Cancer Inst. 2009;101:1538-42. [PMID: 19843888] doi:10.1093/jnci/djp353

63. **Pepe MS, Feng Z, Huang Y, et al.** Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008;167:362-8. [PMID: 17982157]

64. **Fox KA, Poole-Wilson P, Clayton TC, 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;366:914-20. [PMID: 16154018]

65. **Yao X, Gersh BJ, Sangaralingham LR, et al.** Comparison of the CHA_2DS_2-VASc, $CHADS_2$, 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;120:1549-56. [PMID: 28844514] doi:10.1016/j.amjcard.2017.07.051

66. **Hayward RA, Kent DM, Vijan S, et al.** Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol. 2006;6:18. [PMID: 16613605]

67. **Gage BF, Waterman AD, Shannon W, et al.** Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. JAMA. 2001;285:2864-70. [PMID: 11401607]

68. **Gage BF, van Walraven C, Pearce L, et al.** Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. Circulation. 2004;110:2287-92. [PMID: 15477396]

69. **Kent DM, Hayward RA, Griffith JL, 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;113:104-11. [PMID: 12133748]

70. **Thune JJ, Hoefsten DE, Lindholm MG, et al; Danish Multicenter Randomized Study on Fibrinolytic Therapy Versus Acute Coronary Angioplasty in Acute Myocardial Infarction (DANAMI)-2 Investigators.** Simple risk stratification at admission to identify patients with reduced mortality from primary angioplasty. Circulation. 2005;112: 2017-21. [PMID: 16186438]

71. **Yeh RW, Secemsky EA, Kereiakes DJ, et al; DAPT Study Investigators.** Development and validation of a prediction rule for benefit and harm of dual antiplatelet therapy beyond 1 year after percutaneous coronary intervention. JAMA. 2016;315:1735-49. [PMID: 27022822] doi:10.1001/jama.2016.3775

72. **Kernan WN, Viscoli CM, Dearborn JL, et al; Insulin Resistance Intervention After Stroke (IRIS) Trial Investigators.** Targeting pioglitazone hydrochloride therapy after stroke or transient ischemic attack according to pretreatment risk for stroke or myocardial infarction. JAMA Neurol. 2017;74:1319-27. [PMID: 28975241] doi:10.1001/jamaneurol.2017.2136

73. **Knowler WC, Barrett-Connor E, Fowler SE, et al; Diabetes Prevention Program Research Group.** Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med. 2002;346:393-403. [PMID: 11832527]

74. **Kent DM, Vijan S, Hayward RA, et al.** Tissue plasminogen activator was cost-effective compared to streptokinase in only selected patients with acute myocardial infarction. J Clin Epidemiol. 2004;57: 843-52. [PMID: 15485737]

75. **Ioannidis JP, Garber AM.** Individualized cost-effectiveness analysis. PLoS Med. 2011;8:e1001058. [PMID: 21765810] doi:10.1371/journal.pmed.1001058

76. **Schuit E, Li AH, Ioannidis JPA.** How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study. Int J Epidemiol. 2019;48:596-608. [PMID: 30445577] doi:10.1093/ije/dyy239

77. **Kent DM, Jafar TH, Hayward RA, et al.** Progression risk, urinary protein excretion, and treatment effects of angiotensin-converting enzyme inhibitors in nondiabetic kidney disease. J Am Soc Nephrol. 2007;18:1959-65. [PMID: 17475813]

78. **Trikalinos TA, Ioannidis JP.** Predictive modeling and heterogeneity of baseline risk in meta-analysis of individual patient data. J Clin Epidemiol. 2001;54:245-52. [PMID: 11223322]

79. **van Klaveren D, Gönen M, Steyerberg EW, et al.** A new concordance measure for risk prediction models in external validation settings. Stat Med. 2016;35:4136-52. [PMID: 27251001] doi:10.1002/sim.6997

80. **Ioannidis JP, Lau J.** The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol. 1997;50:1089-98. [PMID: 9368516]

81. **Wessler BS, Paulus JK, Lundquist CM, et al.** Tufts PACE Clinical Prediction Model Registry: update 1990 through 2015. Diagn Progn Res. 2017;1:20. doi:10.1186/s41512-017-0021-2

82. **Wessler BS, Lai Yh L, Kramer W, et al.** Clinical prediction models for cardiovascular disease: Tufts Predictive Analytics and Comparative Effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes. 2015;8:368-75. [PMID: 26152680] doi:10.1161/CIRCOUTCOMES.115.001693

83. **Shah ND, Steyerberg EW, Kent DM.** Big data and predictive analytics: recalibrating expectations. JAMA. 2018;320:27-8. [PMID: 29813156] doi:10.1001/jama.2018.5602

84. **Lip GY, Nieuwlaat R, Pisters R, et al.** 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;137:263-72. [PMID: 19762550] doi:10.1378/chest.09-1584

85. **Selker HP, Beshansky JR, Griffith JL, 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. Ann Intern Med. 1998;129:845-55. [PMID: 9867725]

86. **Hess EP, Hollander JE, Schaffer JT, et al.** Shared decision making in patients with low risk chest pain: prospective randomized pragmatic trial. BMJ. 2016;355:i6165. [PMID: 27919865] doi:10.1136/bmj.i6165

87. **Stiell IG, Clement CM, McKnight RD, et al.** The Canadian C-spine rule versus the NEXUS low-risk criteria in patients with trauma. N Engl J Med. 2003;349:2510-8. [PMID: 14695411]

88. **Kuppermann N, Holmes JF, Dayan PS, et al; Pediatric Emergency Care Applied Research Network (PECARN).** Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet. 2009;374:1160-70. [PMID: 19758692] doi:10.1016/S0140-6736(09)61558-0

89. **Wells PS, Anderson DR, Rodger M, 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;135:98-107. [PMID: 11453709]

90. **de Koning L, Merchant AT, Pogue J, et al.** Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: meta-regression analysis of prospective studies. Eur Heart J. 2007;28: 850-6. [PMID: 17403720]

91. **Vazquez G, Duval S, Jacobs DR Jr, et al.** Comparison of body mass index, waist circumference, and waist/hip ratio in predicting incident diabetes: a meta-analysis. Epidemiol Rev. 2007;29:115-28. [PMID: 17494056]

92. **Vickers AJ, Kent DM.** The Lake Wobegon effect: why most patients are at below-average risk. Ann Intern Med. 2015;162:866-7. [PMID: 25867499] doi:10.7326/M14-2767

93. **Farooq V, van Klaveren D, Steyerberg EW, 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;381:639-50. [PMID: 23439103] doi:10.1016/S0140-6736(13)60108-7

94. **Ioannidis JP, Lau J.** Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. Am J Epidemiol. 1998;148:1117-26. [PMID: 9850135]

95. **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. [PMID: 29654822] doi:10.1016/j.jclinepi.2018.04.005

96. **Stern RH.** Individual risk. J Clin Hypertens (Greenwich). 2012;14: 261-4. [PMID: 22458749] doi:10.1111/j.1751-7176.2012.00592.x

97. **Kent DM, Shah ND.** Risk models and patient-centered evidence: should physicians expect one right answer? JAMA. 2012;307: 1585-6. [PMID: 22511683] doi:10.1001/jama.2012.469

98. **Steyerberg EW.** Clinical Prediction Models: A Practical Approach to Development, Valdiation, and Updating. New York: Springer; 2009.

99. **Harrell FE.** Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. New York: Springer; 2015.

100. **Burke JF, Hayward RA, Nelson JP, et al.** Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes. 2014;7:163-9. [PMID: 24425710] doi:10.1161/CIRCOUTCOMES.113.000497

101. **Wang R, Lagakos SW.** Response to letter "More on subgroup analyses in clinical trials." N Engl J Med. 2008;358:2076-7.

102. **Abadie A, Chingos MM, West MR.** Endogenous stratification in randomized experiments. NBER Working Paper no. w19742. Cambridge, MA: National Bureau of Economic Research; 2013. Accessed at http://ssrn.com/abstract=2370198 on 1 May 2019.

103. **Verver D, van Klaveren D, van Akkooi ACJ, 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. [PMID: 29660597] doi:10.1016/j.ejca.2018.02.022

104. **Moons KG, Altman DG, Reitsma JB, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73. [PMID: 25560730] doi:10.7326/M14-0698

105. **Hemingway H, Croft P, Perel P, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595. [PMID: 23386360] doi:10.1136/bmj.e5595

106. **Riley RD, Hayden JA, Steyerberg EW, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10:e1001380. [PMID: 23393429] doi:10.1371/journal.pmed.1001380

107. **Steyerberg EW, Moons KG, van der Windt DA, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10:e1001381. [PMID: 23393430] doi:10.1371/journal.pmed.1001381

108. **Hingorani AD, Windt DA, Riley RD, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 4: stratified medicine research. BMJ. 2013;346:e5793. [PMID: 23386361] doi:10.1136/bmj.e5793

109. **Stewart LA, Clarke M, Rovers M, et al; PRISMA-IPD Development Group.** Preferred Reporting Items for a Systematic review and Meta-Analysis of Individual Participant Data: the PRISMA-IPD state-

ment. JAMA. 2015;313:1657-65. [PMID: 25919529] doi:10.1001/jama.2015.3656
110. **Debray TP, Moons KG, Ahmed I, et al.** A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158-80. [PMID: 23307585] doi:10.1002/sim.5732
111. **Ahmed I, Debray TP, Moons KG, et al.** Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol. 2014;14:3. [PMID: 24397587] doi:10.1186/1471-2288-14-3
112. **Turner RM, Omar RZ, Yang M, et al.** A multilevel model framework for meta-analysis of clinical trials with binary outcomes. Stat Med. 2000;19:3417-32. [PMID: 11122505]
113. **Riley RD, Lambert PC, Abo-Zaid G.** Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221. [PMID: 20139215] doi:10.1136/bmj.c221
114. **Antman EM, Cohen M, Bernink PJ, et al.** The TIMI risk score for unstable angina/non-ST elevation MI: a method for prognostication and therapeutic decision making. JAMA. 2000;284:835-42. [PMID: 10938172]
115. **Johnston SC, Rothwell PM, Nguyen-Huynh MN, et al.** Validation and refinement of scores to predict very early stroke risk after transient ischaemic attack. Lancet. 2007;369:283-92. [PMID: 17258668]
116. **Greenland S.** Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology. 1995;6:450-4. [PMID: 7548361]
117. **Harrell FE.** Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.
118. **Royston P, Sauerbrei W.** Multivariable Model-Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables. Chichester, United Kingdom: J Wiley; 2008.
119. **Kovalchik SA, Varadhan R, Weiss CO.** Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Stat Med. 2013;32:4906-23. [PMID: 23788362] doi:10.1002/sim.5881
120. **Follmann DA, Proschan MA.** A multivariate test of interaction for use in clinical trials. Biometrics. 1999;55:1151-5. [PMID: 11315061]
121. **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. [PMID: 17550609]
122. **van Klaveren D, Vergouwe Y, Farooq V, et al.** Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. J Clin Epidemiol. 2015;68:1366-74. [PMID: 25814403] doi:10.1016/j.jclinepi.2015.02.012
123. **Weinberg CR.** How bad is categorization? [Editorial]. Epidemiology. 1995;6:345-7. [PMID: 7548338]
124. **Bennette C, Vickers A.** Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21. [PMID: 22375553] doi:10.1186/1471-2288-12-21
125. **Schandelmaier S.** Evaluating the Credibility of Effect Modification Claims in Randomized Controlled Trials and Meta-analyses. Hamilton, Ontario, Canada: McMaster Univ; 2019.
126. **Hacke W, Donnan G, Fieschi C, et al; ATLANTIS Trials Investigators.** Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. Lancet. 2004;363:768-74. [PMID: 15016487]
127. **Emberson J, Lees KR, Lyden P, et al; Stroke Thrombolysis Trialists' Collaborative Group.** 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;384:1929-35. [PMID: 25106063] doi:10.1016/S0140-6736(14)60584-5
128. **Bazelier MT, de Vries F, Vestergaard P, et al.** Risk of fracture with thiazolidinediones: an individual patient data meta-analysis. Front Endocrinol (Lausanne). 2013;4:11. doi:10.3389/fendo.2013.00011
129. **Home PD, Pocock SJ, Beck-Nielsen H, et al; RECORD Study Team.** Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet. 2009;373:2125-35. [PMID: 19501900] doi:10.1016/S0140-6736(09)60953-3
130. **Loke YK, Singh S, Furberg CD.** Long-term use of thiazolidinediones and fractures in type 2 diabetes: a meta-analysis. CMAJ. 2009;180:32-9. [PMID: 19073651] doi:10.1503/cmaj.080486
131. **Jafar TH, Stark PC, Schmid CH, et al; AIPRD Study Group.** Proteinuria as a modifiable risk factor for the progression of non-diabetic renal disease. Kidney Int. 2001;60:1131-40. [PMID: 11532109]
132. **Wallach JD, Sullivan PG, Trepanowski JF, et al.** Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med. 2017;177:554-60. [PMID: 28192563] doi:10.1001/jamainternmed.2016.9125
133. **Wallach JD, Sullivan PG, Trepanowski JF, et al.** Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. BMJ. 2016;355:i5826. [PMID: 27884869] doi:10.1136/bmj.i5826
134. **Ioannidis JP.** Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8. [PMID: 18633328] doi:10.1097/EDE.0b013e31818131e7
135. **Paulus JK, Raman G, Rekkas A, et al.** White paper, appendix 1: methods and results of evidence review committee search. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Washington, DC: Patient-Centered Outcomes Research Institute; 2018.
136. **Qian M, Murphy SA.** Performance guarantees for individualized treatment rules. Ann Stat. 2011;39:1180-210. [PMID: 21666835]
137. **Zhang B, Tsiatis AA, Laber EB, et al.** A robust method for estimating optimal treatment regimes. Biometrics. 2012;68:1010-8. [PMID: 22550953] doi:10.1111/j.1541-0420.2012.01763.x
138. **Zhang B, Tsiatis AA, Davidian M, et al.** Estimating optimal treatment regimes from a classification perspective. Stat. 2012;1:103-14. [PMID: 23645940]
139. **Kraemer HC.** Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Stat Med. 2013;32:1964-73. [PMID: 23303653] doi:10.1002/sim.5734
140. **Wallace ML, Frank E, Kraemer HC.** A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. JAMA Psychiatry. 2013;70:1241-7. [PMID: 24048258] doi:10.1001/jamapsychiatry.2013.1960
141. **Tian L, Alizadeh AA, Gentles AJ, et al.** A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2014;109:1517-32. [PMID: 25729117]
142. **Taylor JMG, Cheng W, Foster JC.** Reader reaction to "a robust method for estimating optimal treatment regimes" by Zhang et al. (2012). Biometrics. 2015;71:267-73. [PMID: 25228049] doi:10.1111/biom.12228
143. **Xu Y, Yu M, Zhao YQ, et al.** Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics. 2015;71:645-53. [PMID: 25962845] doi:10.1111/biom.12322
144. **Foster JC, Taylor JM, Kaciroti N, et al.** Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. Biostatistics. 2015;16:368-82. [PMID: 25398774] doi:10.1093/biostatistics/kxu049
145. **Niles AN, Loerinc AG, Krull JL, 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;48:490-500. [PMID: 28577585] doi:10.1016/j.beth.2017.02.001
146. **Petkova E, Tarpey T, Su Z, et al.** Generated effect modifiers (GEM's) in randomized clinical trials. Biostatistics. 2017;18:105-18. [PMID: 27465235] doi:10.1093/biostatistics/kxw035
147. **Cai T, Tian L, Wong PH, et al.** Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011;12:270-82. [PMID: 20876663] doi:10.1093/biostatistics/kxq060

148. **Claggett B, Tian L, Castagno D, et al.** Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. Biostatistics. 2015;16:60-72. [PMID: 25122189] doi:10.1093/biostatistics/kxu037
149. **Basu S, Sussman JB, Rigdon J, et al.** 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:e1002410. [PMID: 29040268] doi:10.1371/journal.pmed.1002410
150. **van Klaveren D, Steyerberg EW, Serruys PW, et al.** The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol. 2018;94:59-68. [PMID: 29132832] doi:10.1016/j.jclinepi.2017.10.021
151. **Iwashyna TJ, Burke JF, Sussman JB, et al.** Implications of heterogeneity of treatment effect for reporting and analysis of randomized trials in critical care. Am J Respir Crit Care Med. 2015;192:1045-51. [PMID: 26177009] doi:10.1164/rccm.201411-2125CP
152. **Groenwold RH, Moons KG, Pajouheshnia R, et al.** Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. J Clin Epidemiol. 2016;78:90-100. [PMID: 27045189] doi:10.1016/j.jclinepi.2016.03.017
153. **Abadie A, Chingos MM, West MR.** Endogenous stratification in randomized experiments. Rev Econ Stat. 2018;100:567-80.
154. **Weisberg HI, Pontes VP.** Post hoc subgroups in clinical trials: anathema or analytics? Clin Trials. 2015;12:357-64. [PMID: 26062595] doi:10.1177/1740774515588096
155. **Zhao L, Tian L, Cai T, et al.** Effectively selecting a target population for a future comparative study. J Am Stat Assoc. 2013;108:527-39. [PMID: 24058223]
156. **Julien M, Hanley JA.** Profile-specific survival estimates: making reports of clinical trials more patient-relevant. Clin Trials. 2008;5:107-15. [PMID: 18375648] doi:10.1177/1740774508089511
157. **Dorresteijn JA, Visseren FL, Ridker PM, et al.** Estimating treatment effects for individual patients based on the results of randomised clinical trials. BMJ. 2011;343:d5888. [PMID: 21968126] doi:10.1136/bmj.d5888
158. **Berger JO, Wang X, Shen L.** A Bayesian approach to subgroup identification. J Biopharm Stat. 2014;24:110-29. [PMID: 24392981] doi:10.1080/10543406.2013.856026
159. **Chen W, Ghosh D, Raghunathan TE, et al.** Bayesian variable selection with joint modeling of categorical and survival outcomes: an application to individualizing chemotherapy treatment in advanced colorectal cancer. Biometrics. 2009;65:1030-40. [PMID: 19210736] doi:10.1111/j.1541-0420.2008.01181.x
160. **Gunter L, Zhu J, Murphy S.** Variable selection for qualitative interactions in personalized medicine while controlling the familywise error rate. J Biopharm Stat. 2011;21:1063-78. [PMID: 22023676] doi:10.1080/10543406.2011.608052
161. **Ternès N, Rotolo F, Heinze G, et al.** Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. Biom J. 2017;59:685-701. [PMID: 27862181] doi:10.1002/bimj.201500234
162. **Djulbegovic B, Ioannidis JPA.** Precision medicine for individual patients should use population group averages and larger, not smaller, groups. Eur J Clin Invest. 2019;49:e13031. [PMID: 30251305] doi:10.1111/eci.13031
163. **Simon R.** Sensitivity, specificity, PPV, and NPV for predictive biomarkers. J Natl Cancer Inst. 2015;107. [PMID: 26109105] doi:10.1093/jnci/djv153
164. **Janes H, Pepe MS, McShane LM, et al.** The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. J Natl Cancer Inst. 2015;107. [PMID: 26109106] doi:10.1093/jnci/djv157
165. **Zhang Z, Nie L, Soon G, et al.** The use of covariates and random effects in evaluating predictive biomarkers under a potential outcome framework. Ann Appl Stat. 2014;8:2336-55. [PMID: 26779295]
166. **Huang Y, Gilbert PB, Janes H.** Assessing treatment-selection markers using a potential outcomes framework. Biometrics. 2012;68:687-96. [PMID: 22299708] doi:10.1111/j.1541-0420.2011.01722.x
167. **Fine JP, Pencina M.** On the quantitative assessment of predictive biomarkers [Editorial]. J Natl Cancer Inst. 2015;107. [PMID: 26109107] doi:10.1093/jnci/djv187
168. **Selker HP, Beshansky JR, Griffith JL; TPI Trial Investigators.** 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;137:87-95. [PMID: 12118963]
169. **Senn S.** Individual response to treatment: is it a valid assumption? BMJ. 2004;329:966-8. [PMID: 15499115]
170. **Song SY, Seo H, Kim G, et al.** Trends in endpoint selection in clinical trials of advanced breast cancer. J Cancer Res Clin Oncol. 2016;142:2403-13. [PMID: 27586374] doi:10.1007/s00432-016-2221-5
171. **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;191:313-9. [PMID: 23715997] doi:10.1007/s00408-013-9479-z
172. **Goldfarb M, Drudi L, Almohammadi M, et al.** Outcome reporting in cardiac surgery trials: systematic review and critical appraisal. J Am Heart Assoc. 2015;4:e002204. [PMID: 26282561] doi:10.1161/JAHA.115.002204
173. **Phillips R, Hazell L, Sauzet O, et al.** Analysis and reporting of adverse events in randomised controlled trials: a review. BMJ Open. 2019;9:e024537. [PMID: 30826796] doi:10.1136/bmjopen-2018-024537
174. **National Research Council.** The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: National Academies Pr; 2010.
175. **Little RJ, D'Agostino R, Cohen ML, et al.** The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367:1355-60. [PMID: 23034025] doi:10.1056/NEJMsr1203730
176. **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use.** Addendum to ICH E9(R1): Statistical Principles for Clinical Trials. Estimands and Sensitivity Analysis in Clinical Trials. 2017.
177. **Hernán MA, Scharfstein D.** Cautions as regulators move to end exclusive reliance on intention to treat. Ann Intern Med. 2018;168:515-6. [PMID: 29554689] doi:10.7326/M17-3354
178. **Ratitch B, Bell J, Mallinckrodt C, et al.** Choosing estimands in clinical trials: putting the ICH E9(R1) into practice. Ther Innov Regul Sci. 2019:2168479019838827. [PMID: 30947539] doi:10.1177/2168479019838827
179. **Mallinckrodt CH, Bell J, Liu G, et al.** Aligning estimators with estimands in clinical trials: putting the ICH E9(R1) guidelines into practice. Ther Innov Regul Sci. 2019:2168479019836979. [PMID: 30955353] doi:10.1177/2168479019836979
180. **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. [PMID: 20442226] doi:10.1136/bmj.c2073
181. **Bond SJ, White IR, Sarah Walker A.** Instrumental variables and interactions in the causal analysis of a complex clinical trial. Stat Med. 2007;26:1473-96. [PMID: 16900567]
182. **Sommer A, Zeger SL.** On estimating efficacy from clinical trials. Stat Med. 1991;10:45-52. [PMID: 2006355]
183. **Hernán MA, Hernández-Díaz S, Robins JM.** Randomized trials analyzed as observational studies. Ann Intern Med. 2013;159:560-2. [PMID: 24018844]
184. **Hernán MA, Robins JM.** Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377:1391-8. [PMID: 28976864] doi:10.1056/NEJMsm1605385
185. **Byar DP.** Assessing apparent treatment–covariate interactions in randomized clinical trials. Stat Med. 1985 Jul-Sep;4:255-63. [PMID: 4059716]
186. **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? [Editorial]. Circ Cardiovasc Qual Outcomes. 2015;8:332-4. [PMID: 26152684] doi:10.1161/CIRCOUTCOMES.115.002038

187. **Decker C, Garavalia L, Garavalia B, et al.** Understanding physician-level barriers to the use of individualized risk estimates in percutaneous coronary intervention. Am Heart J. 2016;178:190-7. [PMID: 27502869] doi:10.1016/j.ahj.2016.03.027

188. **Collins GS, Reitsma JB, Altman DG, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63. [PMID: 25560714] doi:10.7326/M14-0697

189. **Krumholz HM, Ross JS, Gross CP, et al.** A historic moment for open science: the Yale University Open Data Access project and Medtronic [Editorial]. Ann Intern Med. 2013;158:910-1. [PMID: 23778908] doi:10.7326/0003-4819-158-12-201306180-00009

190. **Dahabreh IJ, Kent DM.** Can the learning health care system be educated with observational data? JAMA. 2014;312:129-30. [PMID: 25005647] doi:10.1001/jama.2014.4364

191. **Vickers AJ, Scardino PT.** The clinically-integrated randomized trial: proposed novel method for conducting large trials at low cost [Editorial]. Trials. 2009;10:14. [PMID: 19265515] doi:10.1186/1745-6215-10-14

192. **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;275:562-9. [PMID: 24635449] doi:10.1111/joim.12211

193. **Fiore LD, Lavori PW.** Integrating randomized comparative effectiveness research with patient care. N Engl J Med. 2016;374:2152-8. [PMID: 27248620] doi:10.1056/NEJMra1510057

194. **Franklin JM, Schneeweiss S.** When and how can real world data analyses substitute for randomized controlled trials? Clin Pharmacol Ther. 2017;102:924-33. [PMID: 28836267] doi:10.1002/cpt.857

195. **Byar DP.** Why data bases should not replace randomized clinical trials. Biometrics. 1980;36:337-42. [PMID: 7407321]

196. **Jordan MI, Mitchell TM.** Machine learning: trends, perspectives, and prospects. Science. 2015;349:255-60. [PMID: 26185243] doi:10.1126/science.aaa8415

197. **LeCun Y, Bengio Y, Hinton G.** Deep learning. Nature. 2015;521:436-44. [PMID: 26017442] doi:10.1038/nature14539