Annals of Internal Medicine RESEARCH AND REPORTING METHODS

The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement

David M. Kent, MD, MS; Jessica K. Paulus, ScD; David van Klaveren, PhD; 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

Heterogeneity of treatment effect (HTE) refers to the nonrandom variation in the magnitude or direction of a treatment effect across levels of a covariate, as measured on a selected scale, against a clinical outcome. In randomized controlled trials (RCTs), HTE is typically examined through a subgroup analysis that contrasts effects in groups of patients defined "1 variable at a time" (for example, male vs. female or old vs. young). The authors of this statement present guidance on an alternative approach to HTE analysis, "predictive HTE analysis." The goal of predictive HTE analysis is to provide patient-centered estimates of outcome risks with versus without the intervention, taking into account all relevant patient attributes simultaneously. The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement was developed using a multidisciplinary technical expert panel, targeted literature reviews, simulations to characterize potential problems with predictive approaches, and a deliberative process engaging the expert panel. The authors distinguish 2 categories of predictive HTE approaches: a "risk-modeling" approach, wherein a multivariable model predicts the risk for an outcome and is applied to disaggregate patients within RCTs to define risk-based variation in benefit, and an "effect-modeling" approach, wherein a model is developed on RCT data by incorporating a term for treatment assignment and interactions between treatment and baseline covariates. Both approaches can be used to predict differential absolute treatment effects, the most relevant scale for clinical decision making. The authors developed 4 sets of guidance: criteria to determine when risk-modeling approaches are likely to identify clinically important HTE, methodological aspects of risk-modeling methods, considerations for translation to clinical practice, and considerations and caveats in the use of effect-modeling approaches. The PATH Statement, together with its explanation and elaboration document, may guide future analyses and reporting of RCTs.

Ann Intern Med. 2020;172:35-45. doi:10.7326/M18-3667 **Annals.org**
For author affiliations, see end of text.
This article was published at Annals.org on 12 November 2019.

Medical treatment decisions by clinicians and patients are generally based—implicitly or explicitly—on predictions of outcomes under alternative treatment conditions. Under the paradigm of evidence-based medicine (EBM), the results of randomized controlled trials (RCTs), singly or aggregated in meta-analysis, are the primary evidence used to support these predictions.

Popular approaches to EBM have encouraged the direct application of summary trial results to guide decision making for individuals, as though all patients meeting trial enrollment criteria are likely to similarly experience the benefits and harms of treatment. Yet, recognition has been growing that patients enrolled in trials typically differ in many ways that might be relevant; in particular, they can differ substantially in risk for the outcome and in the balance of benefits and harms of treatment. Thus, there is also growing recognition of the limitations of summary RCT results as tools for prediction in individualized clinical decision making, even among trial-eligible patients (1–3).

Understanding how a treatment's effect can vary across patients (4–10), a concept described as heterogeneity of treatment effect (HTE), is central to the research agenda for both personalized (or precision) medicine and comparative effectiveness research. We define HTE as nonrandom variation in the magnitude or direction of a treatment effect across levels of a covariate (that is, a patient attribute or set of attributes) against a clinical outcome. Because treatment effect can be measured on different scales (such as absolute risk difference or relative risk reduction), HTE is fundamentally a scale-dependent concept (that is, its presence or absence depends on the scale with which the effect is measured) (11).

Extensive literature exists on conventional subgroup analyses, which serially divide the trial population into groups (for example, male vs. female or old vs. young) and examine contrasts in relative treatment effects (12–22). Although potentially useful for exploring hypotheses about factors that modify a treatment effect, these "1-variable-at-a-time" analyses have important limitations. Briefly, low statistical power, multiplicity, and weak prior theory on relative effect modifiers make subgroup analyses prone to both false-negative and false-positive results (23–25). Such analyses also do not provide patient-centered estimates of treatment effects because patients have many attributes that simultaneously affect the outcome of interest and the benefits of treatment. The well-appreciated limitations of subgroup analysis have reinforced the reliance on summary trial results for clinical decision making and the false impression that harm–benefit tradeoffs are similar for all patients meeting trial enrollment criteria (3, 6).

See also:

Editorial comment 63

Web-Only
Related Explanation and Elaboration article

Predictive approaches to HTE analysis are designed to address some of these limitations. The goal of predictive HTE analysis is to provide individualized predictions of treatment effect that are defined by the difference between expected potential outcomes of interest in a particular patient with one intervention versus an alternative and that take into account multiple relevant characteristics simultaneously (7, 26). Although guidance is expanding regarding optimal approaches to prediction modeling (such as PROGRESS [Prognosis Research Strategy] [27, 28] and TRIPOD [Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis] [29]) and some guidance exists for HTE analysis (such as the Patient-Centered Outcomes Research Institute Standards for HTE [30]), no consensus guidelines address specific methodological issues for predictive HTE analysis.

The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement outlines principles, criteria, and key considerations for applying predictive HTE approaches to clinical trials to provide patient-centered evidence in support of decision making. The primary focus of this effort is the identification of clinically important HTE—that is, variation in the risk difference across patient subgroups that may be sufficient to span clinically defined decision thresholds (6, 9, 26). In this article, we summarize criteria to determine when risk-modeling approaches are likely to identify clinically important HTE, review critical methodological aspects of these methods, and identify considerations and caveats for translation to clinical practice. We focus on modeling strategies that use regression analysis but acknowledge a wide set of other approaches (such as tree methods and machine learning). The PATH Statement is intended to be used in conjunction with the explanation and elaboration document (31), which expands on the intent of each recommendation; describes the rationale for each; and discusses related analytic considerations, caveats, and reservations.

PREDICTIVE HTE APPROACHES: RISK MODELING VERSUS EFFECT MODELING

The main goal of predictive HTE analysis is to develop models that can predict which of 2 or more treatments will be better for a particular patient. We distinguish predictive HTE analyses from HTE analyses with other goals, such as those more focused on causal interaction, which include both exploratory (hypothesis-generating) and confirmatory (hypothesis-testing) subgroup analyses. The explanation and elaboration document (31) more fully discusses differing concepts of "interaction" and HTE and contains a glossary of key definitions; a recent narrative review (26), developed as background to the PATH Statement, also provides important context.

Predictive HTE analysis generally comprises 2 steps: variable and model selection to define the reference class (or subgrouping) scheme, and effect estimation across different strata of that scheme. Following a previous review (26), we distinguish 2 distinct approaches to predictive HTE analysis (26). The first is "risk modeling," in which a multivariable model that predicts risk for an outcome is applied to stratify patients within trials to examine risk-based variation in treatment effects. In the 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. Both approaches can be used to predict differential absolute treatment effects—that is, a difference in outcome risks under 2 alternative treatments. We describe each briefly in the following sections.

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.

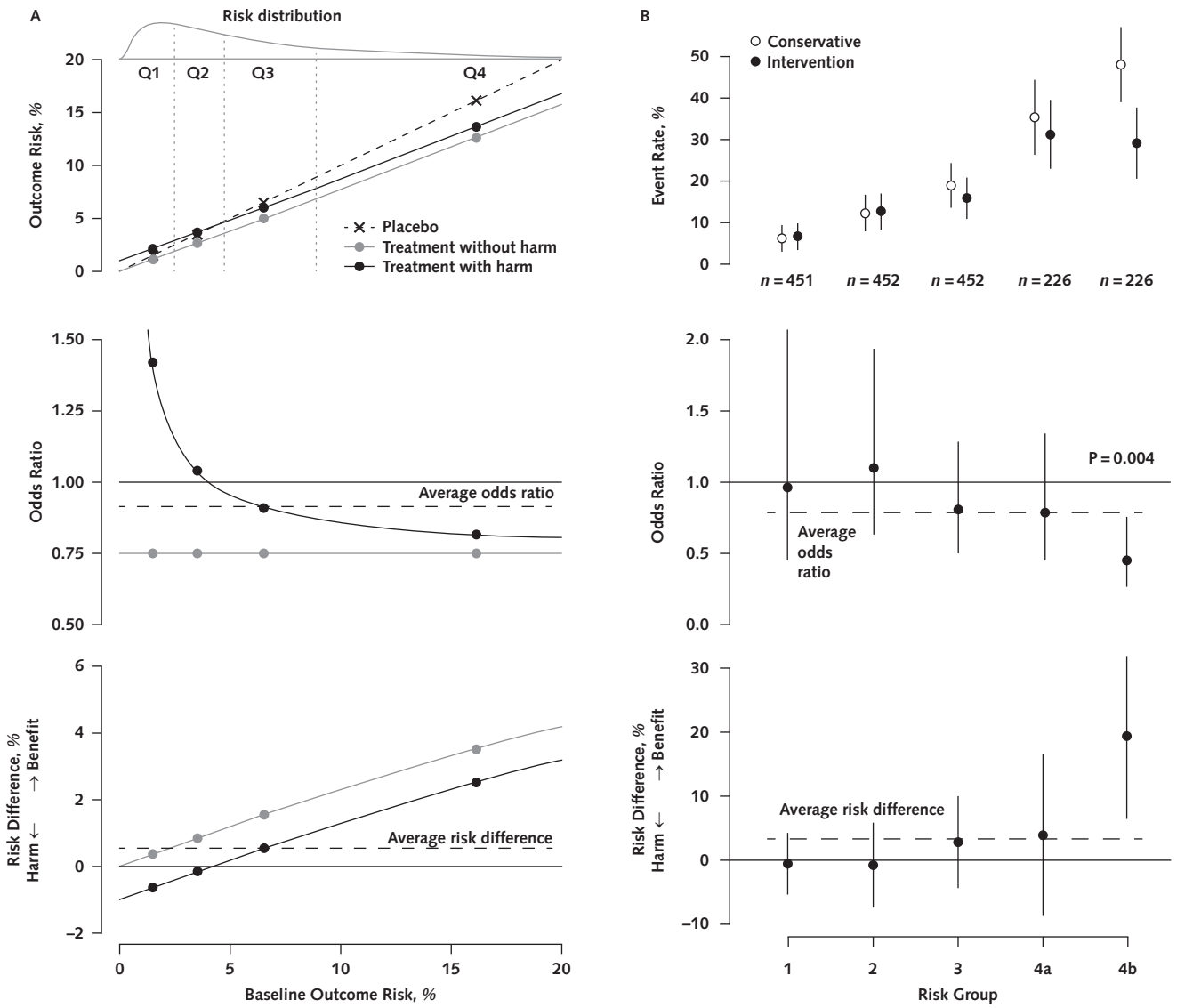
Risk Modeling

Risk modeling relies on the mathematical dependency of treatment effect on the control event rate (Table 1), an observable proxy for outcome risk. When risk is described through a combination of factors (32), the control event rate will typically vary considerably across the trial population. The absolute risk difference—the most clinically important effect measure—will generally vary across risk strata even if the relative risk is the same (32). When a trial population has substantial variation in outcome risk, important differences often exist in harm–benefit tradeoffs (Figure 1) (3, 26). For this reason, risk models can be useful in identifying "clinically important HTE," which is evaluated on the absolute risk difference scale.

A risk-modeling approach is typically done in 2 steps. First, a multivariable regression model that predicts risk for an outcome (usually the primary study outcome) is identified from external sources (an "external model") or developed directly on the trial population without a term for treatment assignment (an "internal model") (equation 1 in Table 2). Next, this model is applied to stratify patients within trials and examine risk-based variation in treatment effects. The essential feature that distinguishes a risk-modeling approach from effect modeling (described in the next section) is that the reference class (or subgrouping) scheme is defined by a model developed "blinded" to treatment assignment.

Figure 1 shows a schematized and an actual example of the risk-modeling approach to trial analysis, in which both relative effects and absolute effects vary by baseline risk across the trial population. In the examples shown, patients are divided using quantiles (such as quartiles) of risk, and treatment effect is estimated in risk groups for reporting purposes. Figure 2 of the explanation and elaboration document (31) shows an alternative form of presentation in which outcomes are shown by predicted risk itself (as a continuous variable), rather than grouped by risk quantiles. Variation in the

Figure 1. 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 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.

B. Stratified results of RITA-3 (Randomized Intervention Trial of unstable Angina 3) (33). 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.

Table 2. Equations Corresponding to Risk-Modeling and Effect-Modeling Approaches

Risk modeling

A multivariable regression model *f* that predicts the risk for an outcome based on risk predictors x_i is identified or developed:

Equation 1: $risk = f(\alpha + \beta_1 \times x_1 + \ldots + \beta_p \times x_p)$

Variation in the treatment effect across risk can be tested statistically on the relative scale through the interaction between a linear predictor of risk ($lp = \beta_1 \times x_1 + \ldots + \beta_p \times x_p$) and treatment assignment *tx*:

Equation 2: $risk = f(\alpha + \beta_{tx} \times tx + \beta_{lp} \times lp + \delta_{lp} \times lp \times tx)$

Including a treatment interaction with the linear predictor of risk permits the relative treatment effect to vary linearly across levels of risk (and permits testing of the statistical significance of this interaction effect, δ_{lp}).

When relative effects across risk strata seem constant, a model with a constant treatment effect may suffice:

Equation 3: $risk = f(\alpha + \beta_{tx} \times tx + \beta_1 \times x_1 + \ldots + \beta_p \times x_p)$,

where the parameter β_{tx} represents a constant risk reduction on the log hazard or log odds scale for treated ($tx = 1$) versus control ($tx = 0$) patients.

Effect modeling

A regression model *f* is developed on RCT data with inclusion of risk predictors x_i, a treatment assignment variable *tx*, and potential treatment interaction terms ($x_i \times tx$):

Equation 4: $risk = f(\alpha + \beta_{tx} \times tx + \beta_1 \times x_1 + \ldots + \beta_p \times x_p + \delta_t \times x_1 \times tx + \ldots + \delta_p \times x_p \times tx)$

RCT = randomized controlled trial.

treatment effect is tested statistically on the relative scale, here using an interaction between the linear predictor of risk and treatment (equation 2 in Table 2). Results are also presented on the scale of absolute risk difference.

In RITA-3 (Randomized Intervention Trial of unstable Angina 3) (Figure 1, *right*), which compared an invasive versus a noninvasive approach for acute coronary syndrome, the eighth of patients at highest risk had an outcome rate approximately 8-fold higher than patients in the lowest-risk quarter. In these high-risk patients, the benefits of therapy outweigh the harms, whereas low-risk patients did not benefit.

For translation to clinical practice, modeling treatment effects across the full risk spectrum (that is, as a continuous variable) can provide more individualized predictions of treatment effect. When relative effects seem to be constant across risk strata, a model with a constant treatment effect may suffice (equation 3 in Table 2). Including a treatment interaction with the linear predictor of risk permits the relative treatment effect to vary linearly across levels of risk (equation 2 in Table 2). Nonlinear interactions between risk and treatment can also be considered. Of note, the exact relationship between risk and treatment effect will depend on the variables included in the risk model.

Effect Modeling

In the second approach, effect modeling, a regression model is developed directly on RCT data with inclusion of risk predictors and a treatment assignment variable and may include treatment interaction terms (equation 4 in Table 2). Because these models include interaction terms between treatment and baseline covariates, they are vulnerable to some of the same problems that undermine conventional subgroup analysis (low power, multiplicity, and limited prior knowledge about important effect modifiers) (26). On rare occasions, highly credible treatment effect interactions exist and should be taken into account when predicting effects in individuals; however, data-driven effect models are prone to bias due to false or exaggerated interaction effects (31, 34), particularly when tests of statistical significance are used to select interaction terms (34, 35). Methods to address these concerns include use of "penalization" to shrink model coefficients to avoid overfitting (36, 37) and use of a 2-step process of first developing the model on randomized data to define the subgrouping (or reference class) scheme and then using a second data set to estimate the treatment effect (38, 39). Practical experience with these data-driven methods is limited, and a robust literature comparing the rapidly emerging approaches is lacking thus far (40–45). Thus, the technical expert panel elected to limit methodological recommendations to those circumstances where specific, highly credible treatment effect interactions have been identified. Although we recognize data-driven effect modeling as a promising area of research, we limit our comments to underscoring caveats and various concerns. A recent review (46) discusses emerging, data-driven approaches to effect modeling.

DEVELOPING THE PATH STATEMENT

To develop the PATH Statement criteria and considerations, we adopted an approach that combined expert opinion, review of the literature, and simulation studies (detailed elsewhere [34]).

The PATH Technical Expert Panel

We assembled a panel of 16 experts who would represent various perspectives on these analyses. The PATH technical expert panel comprised experts in HTE, prediction modeling, clinical trials, and guideline development and a patient advocate (Appendix Table 1, available at Annals.org, provides the full roster).

Literature Review

The cochairs of the technical expert panel (D.M.K. and E.S.) led a literature review of important articles related to the conduct of predictive HTE analyses (26). We developed a library of relevant methodological and applied articles on the topic of predictive HTE analysis in randomized trials, as well as articles related to interaction testing and subgroup analysis. Additional articles were solicited from the panelists and made available to them in an online library. In addition, an evidence review committee (chaired by J.B.W.) did a

systematic scoping review (detailed elsewhere [47]) to identify methodological studies of predictive HTE analysis in RCTs that use regression methods. The goal was to generate an annotated bibliography to guide and support the development of the PATH Statement and to inform future work.

Consensus-Building Process

We used a modified Delphi process (48) to build consensus among panelists to address the main objectives of the PATH Statement. ThinkTank, a cloud-based collaborative platform, facilitated consensus building.

From February to June 2018, the PATH Statement cochairs convened five 2-hour Webinars and one 4-hour in-person meeting to develop consensus. The first meeting was designed to define scope, timeline, and expectations for the PATH Statement. Two additional Webinars were convened in April and July 2019 to consider alterations to recommendations in response to reviewer comments (**Appendix Table 2**, available at Annals.org).

During each Webinar, criteria were scored on a 5-point scale for agreement, importance, and feasibility of assessment. A trained facilitator (J.K.P.) attended Webinars to moderate discussion, structure verbal and electronic communication, and review agreement on criteria. Each vote was recorded, and panelists then had an open discussion centered on areas of disagreement. Items with an SD greater than 1 or mean ratings less than 4 (on 5-point scales) were prioritized for discussion as low-consensus criteria. After discussion, some criteria were deleted, consolidated, or revised, and panelists had the opportunity to revise their judgments with a revote before the next meeting. The criteria for consensus agreement were mean ratings of 4 or more and an SD of 1 or less on the 5-point agreement scale after rounds of discussion and rerating. In the week after a PATH Statement meeting, items reaching consensus and those with limited agreement were distributed to members via e-mail for refinement. **Appendix Tables 3** to **5** (available at Annals.org) provide the results of the final votes on criteria, considerations, and caveats.

Role of the Funding Source

The Patient-Centered Outcomes Research Institute had no role in study design, data collection, analysis, manuscript preparation, or the decision to submit the manuscript for publication.

THE PATH STATEMENT

The PATH Statement comprises 4 sets of guidance (**Figures 2** to **5**) on the conduct of predictive HTE analyses and is briefly summarized in this section; the explanation and elaboration document (31) provides important clarifying details.

First, we developed consensus criteria to define the decision-making, data, design, and analytic context in which the application of risk-modeling approaches is likely to yield clinically informative results (**Figure 2**). The technical expert panel agreed that risk-modeling approaches should generally be applied when an overall treatment effect exists and considered cautiously when it does not, given the potential for false-positive findings. The criteria emphasized conditions where substantial heterogeneity in outcome risk is probable and where this risk heterogeneity is likely to induce important heterogeneity in benefit–harm tradeoffs. This will occur when substantial and identifiable heterogeneity in outcome risk in the trial population is anticipated (that is, a broad case mix) and when treatments are associated with a nontrivial amount of serious harm or burden. **Figure 1** (*right*) shows a clinical example where outcome rates vary dramatically from the high-risk to the low-risk group and a small amount of treatment-related harm may be sufficient to erode much (or all) of the benefit in low-risk patients. Pragmatic and data-related criteria emphasized the availability of several large, randomized, well-conducted clinical trials of contemporary interventions for individual patient meta-analysis and prioritized cases where the variables needed for prediction are routinely available in clinical care. Two additional criteria did not reach consensus (**Appendix Table 3**). Prior work sug-

Figure 2. Consensus criteria: when is a risk-modeling approach to RCT analysis likely to be most valuable?

1. When an overall treatment effect is well established
   A. Subgroup results (including risk-based subgroup results) from overall null trials should be interpreted cautiously
2. When the benefits and harms/burdens of a given intervention are finely balanced (i.e., of similar magnitude on average), increasing the sensitivity of the treatment decision to risk prediction
3. When treatments are associated with a nontrivial amount of serious harm or burden, increasing the importance of careful patient selection
4. When several large, well-conducted RCTs of contemporary interventions are available and appropriate for pooling in individual patient meta-analysis, to provide improved statistical power and broader variation in baseline outcome risk
5. When substantial, identifiable heterogeneity of risk in the trial population is anticipated
   A. When there are validated risk models and well-established risk factors
   B. When case mix heterogeneity is substantial in the trial population
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 (i.e., validation is a high priority)
7. When the clinical variables in the proposed models are routinely available in clinical care

RCT = randomized controlled trial.

Figure 3. Consensus guidance on risk-modeling approaches to identify HTE.

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

Identify or develop a model
3. When available, apply a high-quality, externally developed, compatible risk model to stratify trial results.
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.
5. When developing new risk models or updating externally developed risk models, specify the analytic plan before examining trial data and follow guidance for best practices for prediction model development.

Apply the model and report results
6. Report metrics for model performance for outcome 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.
10. To test the consistency of the relative treatment effect across prognostic risk, a continuous measure of risk (e.g., the logit of risk) may be used in an interaction term with treatment group indicator.

HTE = heterogeneous treatment effects; RCT = randomized controlled trial.

gests that risk modeling might be more informative when the incidence of the outcome is lower (which can lead to more heterogeneity and skewness of risk [32]), but the expert panel was divided on the importance of this criterion.

In Figure 3, we present guidance on best methodological practices to conduct risk modeling to identify HTE, including overarching guidance and that on identifying or developing a risk model, applying the model, and reporting results. This guidance emphasizes the application of a high-quality, externally developed risk model to stratify trial results. Alternatively, when a compatible external model is not available, an internal (or endogenous) risk model that was developed on the entire trial population without a term for treatment assignment (49, 50) can be applied. In either case, the analysis plan should be fully specified before the data are examined. Internal models developed only on the control group (to predict baseline risk, if untreated) are prone to inducing or exaggerating interactions between treatment and risk because of differential fit between the 2 treatment groups. Even slight overfitting on the control group could bias treatment effect estimates across levels of risk (34, 50, 51). Our guidance on reporting the results of risk-stratified analyses underscores the importance of reporting serious treatment-related harms within each stratum to support evaluation of stratum-specific tradeoffs between benefit and harm and the appropriate conduct of statistical hypothesis testing.

Translation of findings from predictive HTE analyses into clinical practice is a complex topic that includes many issues, such as human factors, aspects of model implementation, coping with missing patient characteristics in real-world clinical practice, risk communication, and model transportability. Although a detailed discussion of these challenges is beyond the scope of this project, a set of caveats and considerations for the translation of predictive HTE analyses into clinical practice was developed and is presented in Figure 4. Results should be reported on both absolute and relative scales, but the importance of interpreting treatment effects on the clinically relevant scale (absolute risk difference rather than relative effects) was emphasized (52), as was the importance of external validation and calibration of the risk model used for trial stratification to the target population of interest.

Figure 4. Consensus statements on caveats and considerations before moving to 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.
2. External validation and calibration of risk prediction is important for translation of risk-specific treatment effects into clinical practice.
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.

HTE = heterogeneous treatment effects.

Figure 5. Consensus statements on considerations and caveats in effect modeling for HTE.

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).
2. Avoid 1-variable-at-a-time null hypothesis testing or stepwise selection (e.g., backward selection, forward selection) strategies to select single-variable relative effect modifiers.
3. Avoid the use of regression methods that do not take into account model complexity when estimating coefficients (e.g., "conventional" unpenalized maximum-likelihood regression) when 1 or more treatment-by-covariate interaction terms are included in a treatment effect model.
4. Avoid evaluating models that predict treatment benefit using only conventional metrics for outcome risk prediction (e.g., based on discrimination and calibration of outcome risk prediction).

HTE = heterogeneous treatment effects.

When highly credible relative effect modifiers have been identified, they should be incorporated into prediction models using multiplicative treatment-by-covariate interaction terms (Figure 5). Credibility should be evaluated using rigorous multidimensional criteria (such as those described by the Instrument for assessing the Credibility of Effect Modification ANalyses [ICEMAN tool] [53] and including the strength of a priori direct or indirect evidence) and should not rely solely on statistical criteria (such as *P* value thresholds). We anticipate that highly credible interactions will generally be rare when randomized clinical trials of conventional size are used. In the absence of such interactions, the technical expert panel advised caution in the use of data-driven effect models for the following reasons: Practical experience with data-driven effect modeling (that is, where statistical approaches are used to explore and select out effect modifiers on the relative scale) is limited, these flexible approaches are vulnerable to false discoveries of promising subgroup effects, the variety of competing approaches is growing, and robust literature comparing the different approaches is lacking. Thus, we restricted our recommendations to a set of considerations and caveats for these more "aggressive" modeling techniques (that is, those using more degrees of freedom for estimation of the differential treatment effect). These statements emphasized the need to address the potential for overfitting that can lead to signals of treatment benefit and harm even for interventions that are completely ineffective and innocuous (34). This is a particular problem for modeling RCT data, where power for statistical interaction of relative effects is generally limited, as are opportunities for external validation.

The PATH Explanation and Elaboration Document

A supporting explanation and elaboration document (31), available at Annals.org, clarifies and expands on the motivation and reservations regarding items in the PATH Statement. The document explains each recommendation in more detail and provides (where relevant) clinical applications of methods, supporting methodological evidence, and caveats or limitations. It also describes special considerations for evaluating models that predict benefit. The literature reviews done as part of this project were used to justify the rationale for guidance statements or criteria. The explanation and elaboration document was developed after the PATH Statement criteria reached final consensus, with input from panelists via several face-to-face meetings, teleconferences, and iterations among the authors. Additional revisions were made after the document was shared with the whole PATH Statement group before final approval.

Discussion

The field of EBM has historically emphasized the results of randomized trials (and their meta-analysis) as the best evidence for clinical decision making. However, patient attributes may influence the probability of the outcome of interest and the benefits of treatment. The goal of personalized EBM can be conceived as the identification of an optimal subgrouping (or reference class) scheme, based on all relevant patient characteristics, that yields more individualized estimates of treatment effects for each patient than the average trial result, thus improving overall outcomes (26). In fact, the most common definition of EBM from more than 20 years ago anticipates the need to make decisions for individual patients (54). Thus, the goal of personalization has been at the core of EBM since its inception, although the limitations of summary trial results in supporting this goal have been inconsistently recognized.

The PATH Statement was developed to address this gap. Although the statement and its explanation and elaboration document are supported by a substantial and growing body of evidence, and although they build on prior efforts to offer methodological guidance on predictive approaches to HTE (5, 6), this represents the first such guidance developed by a diverse set of experts and stakeholders with differing views and perspectives that involved an iterative process of discussion, feedback, and revision. The guidance thus aims to assist a diverse set of relevant stakeholders, including researchers, regulators, industry professionals, and guideline-writing bodies.

We focused on risk modeling, and we acknowledge that the more comprehensive and flexible effect-modeling approach (incorporating treatment interaction terms with individual effect modifiers) holds promise—particularly in data sets that are substantially larger than those from conventional RCTs. Capturing the benefits of effect modeling while avoiding the potential harms of overfitting is an area of intense research interest (36, 38, 39, 49, 55–61) and a central challenge for future study. Nevertheless, considerable evidence suggests that risk-modeling approaches can frequently provide clinically important insights—beyond those provided by overall trial results—that can directly improve decision making (33, 40–45, 62, 63).

The PATH Statement focuses on regression-based prediction in randomized trials. A broad and evolving tool kit exists for data-driven approaches to predicting patient benefit, including machine-learning techniques (46, 64). As experience grows, we anticipate stronger methodological and evidentiary guidance that will allow a fuller understanding of the appropriate contexts and the advantages and limitations of these more flexible modeling methods. In addition, although observational studies seem to offer many important advantages over conventional RCTs for more refined analyses (that is, enhanced statistical power and patient heterogeneity), when these data may be sufficiently debiased for reliable determination of treatment effects or HTE is not well understood (9, 65). Our PATH Statement identified these issues as high research priorities.

Predictive HTE methods can often be usefully applied to individual large clinical trials (33, 40–45, 62, 63). However, the PATH Statement authors recognize that fully realizing the goals of improved evidence personalization also depends on increased collaborative efforts to create pooled data substrates that are more conducive to these analyses than individual trials, and on the implementation of innovative trial designs—including those sampling larger and broader populations—that may enrich the heterogeneity of clinical trial populations. The PATH Statement is intended to encourage and motivate these innovations.

We also need research to better integrate clinical prediction into practice (66, 67); to understand how to individualize clinical practice guidelines; to establish or extend reporting guidelines (29); to establish new models of data ownership to facilitate data pooling (68); and to reengineer the clinical research infrastructure to support substantially larger, clinically integrated trials that are sufficiently powered to determine HTE (69). Many recent and ongoing organizational and technical advances should enable this evolution (68, 70–73). The collaborative work in the field of genetic epidemiology (74) may serve as a useful model for HTE prediction if we are to optimally address many of the challenges to individualizing evidence.

The PATH technical expert panel recognizes the inherent difficulties and fundamental limitations of using group data to estimate treatment effects in individuals (9). In particular, individual treatment effects are inherently unobservable (in parallel-group studies), and individual patients do not have uniquely identifiable risks (9, 26, 75–77). Nevertheless, clinicians are required to make decisions 1 patient at a time. The PATH Statement provides guidance for analytic approaches that seek to advance our ability to provide more patient-centered estimates of treatment effects. We present it as an important formative step in a long-term research effort to better personalize evidence from comparative effectiveness data.

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.); 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 the figures and supporting appendix tables. 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. Hayward reports grants from the National Institute of Diabetes and Digestive and Kidney Diseases and the Veterans Affairs Health Services Research and

Development Service during the conduct of the study. 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=M18-3667.

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 and author contributions are available at Annals.org.

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. **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]
5. **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]
6. **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
7. **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
8. **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
9. **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
10. **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
11. **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.
12. **Rothwell PM.** Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176-86. [PMID: 15639301]
13. **Lagakos SW.** The challenge of subgroup analyses–reporting without distorting. N Engl J Med. 2006;354:1667-9. [PMID: 16625007]
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, 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
27. **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
28. **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
29. **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
30. **Patient-Centered Outcomes Research Institute Methodology Committee.** The PCORI Methodology Report. 2019.
31. **Kent DM, van Klaveren D, Paulus JK, et al.** The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement: explanation and elaboration. Ann Intern Med. 2020;172:W1-25. doi:10.7326/M18-3668

32. **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
33. **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]
34. **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
35. **Ioannidis JP.** Why most discovered true associations are inflated. Epidemiology. 2008;19:640-8. [PMID: 18633328] doi:10.1097/EDE.0b013e31818131e7
36. **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
37. **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
38. **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
39. **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
40. **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]
41. **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]
42. **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]
43. **Yusuf S, Diener HC, Sacco RL, et al; PRoFESS Study Group.** Telmisartan to prevent recurrent stroke and cardiovascular events. N Engl J Med. 2008;359:1225-37. [PMID: 18753639] doi:10.1056/NEJMoa0804593
44. **Kovalchik SA, Tammemagi M, Berg CD, et al.** Targeting of low-dose CT screening according to the risk of lung-cancer death. N Engl J Med. 2013;369:245-54. [PMID: 23863051] doi:10.1056/NEJMoa1301851
45. **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
46. **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
47. **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.
48. **Murphy MK, Black NA, Lamping DL, et al.** Consensus development methods, and their use in clinical guideline development. Health Technol Assess. 1998;2:i-iv, 1-88. [PMID: 9561895]
49. **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
50. **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
51. **Abadie A, Chingos MM, West MR.** Endogenous stratification in randomized experients. Cambridge, MA: National Bureau of Economic Research; 2013. Accessed at http://ssrn.com/abstract=2370198 on 16 December 2018.
52. **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
53. **Schandelmaier S.** Evaluating the Credibility of Effect Modification Claims in Randomized Controlled Trials and Meta-analyses. Hamilton, Ontario, Canada: McMaster Univ; 2019.
54. **Sackett DL, Rosenberg WM, Gray JA, et al.** Evidence based medicine: what it is and what it isn't [Editorial]. BMJ. 1996;312:71-2. [PMID: 8555924]
55. **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]
56. **Foster JC, Taylor JM, Ruberg SJ.** Subgroup identification from randomized clinical trial data. Stat Med. 2011;30:2867-80. [PMID: 21815180] doi:10.1002/sim.4322
57. **Chen G, Zhong H, Belousov A, et al.** A PRIM approach to predictive-signature development for patient stratification. Stat Med. 2015;34:317-42. [PMID: 25345685] doi:10.1002/sim.6343
58. **Wang R, Schoenfeld DA, Hoeppner B, et al.** Detecting treatment-covariate interactions using permutation methods. Stat Med. 2015;34:2035-47. [PMID: 25736915] doi:10.1002/sim.6457
59. **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
60. **Künzel SR, Sekhon JS, Bickel PJ, et al.** Meta-learners for estimating heterogeneous treatment effects using machine learning. 2018. Accessed at https://arxiv.org/abs/1706.03461 on 16 December 2018.
61. **Schuler A, Baiocchi M, Tibshirani R, et al.** A comparison of methods for model selection when estimating individual treatment effects. 2018. Accessed at https://arxiv.org/abs/1804.05146v2 on 16 December 2018.
62. **Califf RM, Woodlief LH, Harrell FE Jr, et al; GUSTO-I Investigators.** Selection of thrombolytic therapy for individual patients: development of a clinical model. Am Heart J. 1997;133:630-9. [PMID: 9200390]
63. **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
64. **Jordan MI, Mitchell TM.** Machine learning: trends, perspectives, and prospects. Science. 2015;349:255-60. [PMID: 26185243] doi:10.1126/science.aaa8415
65. **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
66. **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
67. **Selker HP.** Systems for comparing actual and predicted mortality rates: characteristics to promote cooperation in improving hospital care [Editorial]. Ann Intern Med. 1993;118:820-2. [PMID: 8470856]
68. **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
69. **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
70. **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
71. **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
72. **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
73. **Institute of Medicine.** Redesigning the Clinical Effectiveness Research Paradigm: Innovation and Practice-Based Approaches: Workshop Summary. Washington, DC: National Academies Pr; 2010.
74. **Ioannidis JP, Loy EY, Poulton R, et al.** Researching genetic versus nongenetic determinants of disease: a comparison and proposed unification. Sci Transl Med. 2009;1:7ps8. [PMID: 20368180] doi:10.1126/scitranslmed.3000247
75. **Goodman SN.** Probability at the bedside: the knowing of chances or the chances of knowing? [Editorial]. Ann Intern Med. 1999;130:604-6. [PMID: 10189332]
76. **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
77. **Stern RH.** Individual risk. J Clin Hypertens (Greenwich). 2012;14: 261-4. [PMID: 22458749] doi:10.1111/j.1751-7176.2012.00592.x

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

Appendix Table 1. PATH Technical Expert Panel

Name	Affiliation	Expertise
David Kent, MD, MS*	Tufts Medical Center	Predictive modeling and comparative effectiveness research
Ewout Steyerberg, PhD*	Leiden University Medical Center	Predictive modeling and medical decision making
Naomi Aronson, PhD	Blue Cross and Blue Shield Association; Patient-Centered Outcomes Research Institute	Dissemination and implementation science, regulatory science, and methodological standards
Ralph D'Agostino, PhD	Boston University	Cardiovascular risk prediction models, biostatistics, and epidemiology
Steven Goodman, MD, MHS, PhD	Stanford University	Measurement and synthesis of research evidence, Bayesian quantitation, and qualitative approaches
Rodney Hayward, MD	University of Michigan	Quality measurement and improvement for chronic diseases
John P.A. Ioannidis, MD, DSc	Stanford University	Evidence-based medicine, clinical and molecular epidemiology, research methods, statistics, and genomics
Bray Patrick-Lake, MFS	Duke University	Patient advocacy
Sally Morton, PhD	Virginia Tech	Biostatistics, evidence synthesis, and pragmatic studies
Sharon-Lise Normand, PhD	Harvard Medical School	Statistical methods for health services and regulatory policy research
Michael Pencina, PhD	Duke University	Biostatistics and informatics
Joseph Ross, MD	Yale University	Health services research methods and open data sources
Harry Selker, MD, MSPH	Tufts Medical Center	Clinical study design, data analysis, and computer-based mathematical models
Ravi Varadhan, PhD	Johns Hopkins University	Computational statistics, patient-centered outcomes research, aging, and frailty
Andrew Vickers, PhD	Memorial Sloan Kettering	Research methodology, randomized trials, surgical outcomes research, and molecular marker studies
John B. Wong, MD	Tufts Medical Center	Clinical decision making, cost-effectiveness, and health technology assessment

PATH = Predictive Approaches to Treatment effect Heterogeneity.
* Technical expert panel cochair.

Appendix Table 2. TEP Votes on PATH Statement Revisions Held 1 May 2019 and 17 July 2019 (7 Voters)

Two additional meetings of the TEP were held to consider revisions to the PATH Statement criteria, caveats, and considerations in response to editorial and reviewer comments. At the July meeting, all TEP members present (7 members) voted unanimously to include the following new statement related to the inclusion of highly credible effect modifiers in prediction models in Figure 5.

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).

PATH = Predictive Approaches to Treatment effect Heterogeneity; TEP = technical expert panel.

Appendix Table 3. Results of the Final TEP Vote on Criteria to Identify When a Risk-Modeling Approach to RCT Analysis Is Likely to Be of Most Value (11 Voters)*

Criteria	Agree/ Disagree		Importance		Feasibility of Assessment	
	Mean	SD	Mean	SD	Mean	SD
Included						
1. When the decision threshold is near the population average, or (for trials) where the benefits and harms/costs of a given therapy are finely balanced on average	4.82	0.57	4.55	0.89	3.91	1.00
2. When treatments are associated with even a small amount of serious treatment-related harm	3.91	1.08	3.64	1.07	3.36	1.07
3. When several large RCTs of contemporary therapies are available and appropriate for pooling (to provide a sufficiently powered and unconfounded treatment comparison of relevant therapies/cotreatments)	4.27	0.75	3.82	0.57	4.91	0.29
4. When the clinical field is mature and there are validated risk models and well-established risk factors	4.36	0.64	3.82	1.03	4.45	0.66
5. When there is substantial case mix heterogeneity in the trial population	4.55	0.50	4.64	0.48	3.36	0.98
6. When promising treatment selection models have been previously developed on clinical trial data (i.e., validation is a high priority)	4.45	0.66	4.18	0.72	4.27	0.62
7. When the clinical variables in the proposed models are routinely available in clinical care	4.55	0.50	4.18	0.94	4.27	0.86
Excluded						
When the outcome rate is lower	2.91	1.24	2.55	0.78	4.00	1.13
When the 2 treatment groups are very clinically different (e.g., medicine vs. surgery)	2.73	1.05	3.09	0.67	4.36	0.98

RCT = randomized controlled trial; TEP = technical expert panel.
* Values are numbers.

Appendix Table 4. Results of Final Consensus Vote on Risk Modeling Guidance to Identify HTE (13 Voters)*

Guidance Statement	Agree/ Disagree		Importance	
	Mean	SD	Mean	SD
1. Reporting RCT results stratified by a risk model is encouraged when overall trial results are positive.	4.31	0.82	3.85	1.29
2. When available, apply a well-accepted, externally developed, RCT-compatible risk model to stratify trial results.	4.31	0.72	3.85	1.23
3. Consider using observational data for model development.	3.77	0.89	3.23	1.19
3.1. The eligibility criteria for the observational cohort should closely align with those in the trial or be even broader; predictor and outcome variable definitions should be very similar to those available in the RCT.	3.92	0.62	3.62	1.21
4. Consider developing a model to stratify trial results using the entire trial population (blinded to treatment assignment); avoid modeling on the control group only.	4.23	0.58	3.46	1.22
5. When developing new models, follow guidance for best practices for prediction model development (see TRIPOD Statement explanation and elaboration [78]).	4.15	1.29	3.69	1.64
6. Report metrics for model performance on the RCT, including measures of discrimination and calibration (when appropriate).	4.08	1.33	3.62	1.64
7. Report distribution of predicted risk (or risk score) in each group of the trial and in the overall study population.	3.85	1.23	3.38	1.50
7.1. Risk reporting should allow readers to assess the full distribution of risk in the study population either graphically (e.g., histograms or box-and-whisker plots) or by including information on the mean, SD, median, and interquantile range.	3.77	1.19	3.31	1.49
7.2. Report how relative and absolute risk reduction varies in a risk-stratified analysis.	4.00	1.30	3.46	1.60
8. Null hypothesis statistical testing for risk-based HTE (e.g., the *P* value for the interaction of treatment with the linear predictor of risk) may be performed and reported but has a limited role in the clinical interpretation of results.	3.54	1.34	2.77	1.42
9. Consider external validation and calibration of outcome risk prediction models before clinical dissemination or implementation.	4.38	0.84	3.62	1.33
10. Report results in clinically relevant terms (e.g., number of patients treated, number of events avoided with and without use of the model) and consider decision analytic approaches.	4.23	0.58	3.85	1.35

HTE = heterogeneity of treatment effect; RCT = randomized controlled trial; TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.
* Values are numbers.

Appendix Table 5. Results of Final TEP Vote on Caveats of Treatment Effect–Modeling Approaches to Identify HTE (9 voters)*

Caveats	Agree/ Disagree		Importance	
	Mean	SD	Mean	SD
1. When selecting individual variable relative effect modifiers in a treatment effect model, avoid 1-variable-at-a-time null hypothesis testing or stepwise selection (e.g., backward selection, forward selection) strategies to select individual variable relative effect modifiers.	4.78	0.42	4.44	0.68
2. Avoid the use of regression methods that do not take into account model complexity when estimating coefficients (e.g., unpenalized maximum likelihood, least-squares regression) when ≥1 treatment interaction term is included in a treatment effect model.	4.33	0.47	4.33	0.47
3. Avoid evaluating treatment effect models using only conventional metrics for outcome prediction (e.g., based only on discrimination and calibration of outcome risk prediction).	4.56	0.50	4.67	0.47

HTE = heterogeneity of treatment effects; TEP = technical expert panel.
* Values are numbers.

Web-Only Reference

78. **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