Impact of the PATH Statement on Analysis and Reporting of Heterogeneity of Treatment Effect in Clinical Trials: A Scoping Review

Running Title: Predictive Modeling of Individualized Treatment Effects in RCTs

Joe V Selby [*]

Carolien C H M Maas [†,‡]

Bruce H Fireman [§]

David M Kent [†]

* Division of Research, Kaiser Permanente Northern California, Oakland, CA (emeritus)

[†] Tufts Predictive Analytics and Comparative Effectiveness Center, Tufts University School of Medicine, Boston MA

[‡] Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands

[§] Division of Research, Kaiser Permanente Northern California, Oakland, CA

Corresponding Author: Joe V Selby MD, MPH
3101 Benvenue Avenue
Berkeley, CA 94705
jvselby@outlook.com

Word Count: 4116

ABSTRACT (295 words)

Background: The Predictive Approaches to Treatment Effect Heterogeneity (PATH) Statement provides guidance for using predictive modeling to identify differences (i.e., heterogeneity) in treatment effects (benefits and harms) among participants in randomized clinical trials (RCTs). It distinguished risk modeling, which uses a multivariable model to predict risk of trial outcome(s) and then examines treatment effects within strata of predicted risk, from effect modeling, which predicts trial outcomes using models that include treatment, individual participant characteristics and interactions of treatment with selected characteristics.

Purpose: To describe studies of heterogeneous treatment effects (HTE) that use predictive modeling in RCT data and cite the PATH Statement,

Data Sources: The Cited By functions in PubMed, Google Scholar, Web of Science and SCOPUS databases (Jan 7, 2020 - June 5, 2023).

Study Selection: 42 reports presenting 45 predictive models.

Data Extraction: Double review with adjudication to identify risk and effect modeling and examine consistency with Statement consensus statements. Credibility of HTE findings was assessed using criteria adapted from the Instrument to assess Credibility of Effect Modification Analyses (ICEMAN). Clinical importance of credible HTE findings was also assessed.

Data Synthesis: The numbers of reports, especially risk modeling reports, increased year-on-year. Consistency with consensus statements was high, except for two: only 15 of 32 studies with positive overall findings included a risk model; and most effect models explored many candidate covariates with little prior evidence for effect modification. Risk modeling was more likely than effect modeling to identify both credible HTE (14/19 vs 5/26) and clinically important HTE (10/19 vs 4/26).

Limitations: Risk of reviewer bias: reviewers assessing credibility and clinical importance were not blinded to adherence to PATH recommendations.

Conclusions: The PATH Statement appears to be influencing research practice. Risk modeling often uncovered clinically important HTE; effect modeling was more often exploratory.

INTRODUCTION

Findings from randomized controlled trials (RCTs) have limitations for patients making personal treatment decisions. (1-6) RCTs are usually planned and sized to estimate overall or average treatment effects in trial populations. Even in very positive RCTs, some patients do not benefit from the study treatment, and some experience adverse effects of treatment. Patients have many characteristics that might influence their own likelihood or "risk" for study outcomes or for experiencing either benefit or adverse effects from the treatment.

Until recently, guidance for identifying possible differences, or heterogeneity, of treatment effects (HTE) among RCT participants (1,7-10) has focused on considering one characteristic at a time, testing hypotheses for "effect modification" or statistical interaction of the characteristic with treatment. Analyses test whether treatment effects differ between patient subgroups, such as men vs. women or persons with vs. without diabetes. Relative treatment effects (risk, odds, or hazard ratios) are usually compared. These subgroup comparisons are almost always underpowered for detecting interactions in RCTs populations, leading to false negative findings. At the same time, if many subgroups with low prior probabilities for effect modification are explored, chances for false positive findings are also high. Thus, guidelines have consistently recommended limiting the number of subgroups examined, ideally to those with prior evidence or strong biologic or clinical rationale for HTE and using caution in interpreting apparent interactions or applying findings to clinical practice. A fundamental limitation of "one-variable-at-a-time" subgroup analyses is that they do not provide a single treatment effect prediction for an individual because individuals simultaneously belong to multiple subgroups that can vary in whether or how they appear to benefit.

Both patient-centered outcomes research (11) and precision medicine (12) have heightened interest in identifying important HTE to improve individualized clinical decision-making. The authorizing legislation (13) for the Patient-Centered Outcomes Research Institute (PCORI)

explicitly instructed PCORI to require awardees to look for individual differences in the effectiveness of health care treatments and services.

In 2018, an expert panel funded by PCORI described a new approach to HTE analyses. The Predictive Approaches to Treatment Heterogeneity (PATH) Statement (14,15), proposed predictive modeling for identifying clinically important HTE. The Predictive Approaches to Treatment Heterogeneity (PATH) Statement (14,15), proposed predictive modeling for identifying clinically important HTE. Predictive modeling accounts for the effects of multiple patient attributes (independent variables) on trial outcomes simultaneously and can produce individualized predictions of potential benefits and risks of study treatments .The Statement emphasized that important HTE can be seen on the absolute as well as relative scales (i.e., as variation among patients in treatment-related *differences* in outcomes as well as ratios) and that absolute treatment effects are more useful than relative effects for clinical decision-making.

The PATH Statement distinguished two approaches to predictive modeling. “Risk modeling” focuses on one potential effect modifier: the individual’s risk (or probability) of experiencing an outcome (usually the trial’s primary outcome). (16) Multiple baseline characteristics are incorporated into models predicting risk for the outcome and a risk score is generated for each participant. In a second step, treatment effects are examined across strata (e.g. by quartiles) of predicted risk. The Statement encouraged use of previously developed, validated prediction models when available and appropriate to the RCT population, but also suggested that if an appropriate model is not available, one can be developed internally, using observed study outcomes, covariates measured at baseline and all participants regardless of treatment assignment.

Risk modeling has both a mathematical and an empirical rationale in evaluating possible HTE. It builds on the clinical intuition that patients at greater risk for study outcome(s) have more to gain

from a beneficial treatment. Mathematically, baseline risk is linked to the absolute treatment effect as follows:

$$ARR = CER\ (1\text{-}RR) \qquad (1)$$

For any value of relative treatment effect, or relative risk reduction (1 – RR), the absolute risk reduction (ARR) increases as baseline risk (the control event rate, CER) increases. This relationship has been called "risk magnification," (17) but is more accurately a "benefit magnification" when the overall treatment effect is beneficial. Implicit in Equation 1 and the concept of benefit magnification is the assumption that the relative treatment effect (1-RR) is constant across baseline risk. Clinical guidelines and disease management strategies also make this assumption in recommending that persons at greater risk be treated first or more aggressively. In re-analyses of 14 RCTs with positive results, (18) 13 suggested benefit magnification, but in the 14th, (19), a significant interaction of baseline risk and treatment effect was found. Risk modeling allows for testing this assumption by modeling study outcomes as a function of the risk score, treatment assignment, and a statistical interaction between the two. Findings that relative treatment effects vary significantly across baseline risk could markedly change individual treatment recommendations.

In the second approach, "effect modeling", a model predicting the trial's outcome is developed within the RCT data and includes independent variables for treatment assignment, individual patient characteristics and interactions of treatment with selected characteristics. This allows direct estimation of the predicted treatment effect for individual participants. The Statement recognized the more exploratory or "data-driven" nature of most effect modeling and extended the general cautions of earlier guidelines, urging that only candidate treatment interaction terms with strong pre-existing evidence for HTE be included. It recommended use of statistical methods that penalize or "shrink" coefficient estimates to correct for over-fitting and urged caution in interpreting findings. The Statement also recognized the emergence and potential

importance of newer data-driven machine-learning approaches to effect modeling for exploring more complex multi-level interactions among multiple variables but suggested that this field is not yet mature enough to justify specific recommendations or approaches.

This review evaluates the Statement's impact following the standard methodology for scoping reviews (20). We review reports that have appeared since its publication, cited it and presented predictive models of potential HTE using RCT data. We assess consistency of analyses with the Statement's Consensus Criteria (Supplement, Boxes A, B, D) and determine whether authors claimed that HTE was present, on either relative or absolute scales. We adapted criteria of the Instrument to assess Credibility of Effect Modification Analyses (ICEMAN) (21) to determine whether claims were credible and, if so, whether the HTE appeared to be clinically important, which the Statement defined as variation across patients in the treatment effect sufficient to span clinically defined decision thresholds, supporting differing treatment recommendations for patient subgroups.

METHODS:

Identification of Reports for Inclusion. Using the Cited By functions in PubMed, Google Scholar, Web of Science and the SCOPUS database, we sought reports appearing after the Statement's publication (January 7, 2020) through June 5, 2023 that presented analyses or re-analyses of data from RCTs using multivariable predictive modeling to identify HTE. Restricting the search to articles citing the Statement allows the assumption that authors were aware of its concepts and recommendations. The sources other than PubMed allowed inclusion of non-peer-reviewed reports posted on pre-print archives as well as dissertations posted on institutional websites.

A total of 211 citations were identified (Figure 1). Fifty-eight (22-79) involved analysis or re-analysis of data from one or more RCTs. In three instances, we combined two publications from

the same authors and trial(s) (35-40). Thirteen reports (22-34) were excluded because they did not present a predictive model as defined by the Statement. Reasons for exclusion are presented in Supplement Table S1.

Review of Predictive Model Reports. Variables collected during review and coding instructions are presented in Supplemental Tables S2 and S3, respectively. Features describing source RCT(s) were collected by the lead author (JS)., including 4 features the Statement suggested make risk modeling particularly likely to be of value for identifying clinically important HTE (Box A and Table S4 in Supplement). All aspects of analyses and findings for HTE were doubly reviewed by the lead author and one co-author (BF,DK,CM). An initial "learning" set of six reports was reviewed, discussed and resolved by all study co-authors. Thereafter, co-reviewers discussed and resolved initial disagreements.

Review determined whether a report used risk modeling, effect modeling or both. Effect models were further classified into those using regression methods (e.g., ordinary least squares, logistic, proportional hazards) and those that used more flexible, non-parametric data-driven machine-learning algorithms (e.g., 80-83).

Consistency with PATH Statement Criteria and Considerations. The Statement offered 10 guidance criteria for risk modeling (Supplement, Box B). The first recommended that risk modeling be conducted whenever an RCT had a positive overall result. Seven additional criteria were assessed, including using an external risk score (if available), including both treatment arms rather than only the control arm if developing the risk model internally; pre-specifying the analytic plan, including cut points for subgrouping of risk scores, before analyses; reporting risk model performance metrics (i.e., discrimination and calibration) when applied to the RCT population; presenting risk score distributions separately by treatment arm; reporting both absolute and relative effect sizes when reporting risk-stratified treatment effects; and reporting

important adverse treatment effects by risk stratum when present. One remaining criterion could not be assessed at review and one was not a clearcut recommendation (see Supplement, Box B). For effect models, consistency with four Statement consensus criteria (Supplement, Box D) was recorded. These included limiting model covariates to those with strong prior evidence for effect modification; taking steps to reduce risks of model overfitting by applying penalization/regularization procedures to model coefficients (e.g., least absolute shrinkage and election operator (LASSO) (84), penalized ridge regression (85), elastic net regularization (86)) and/or using internal cross-validation; validating final model performance in a dataset external to the population in which it was developed (credit was given for using either an entirely distinct RCT dataset or a non-random subset of the original population selected before analyses on the basis of either geography (e.g., trial sites) or time of enrollment); and not relying solely on metrics intended for evaluating risk prediction when evaluating the performance of treatment effect models. In addition, we noted whether authors evaluated model performance in terms of predicting patient-specific treatment effects, including use of recently developed performance metrics for this purpose (e.g., 87-89).

Assessment of Credible and Clinically Important HTE. For reports that claimed to have identified HTE on either the absolute or relative scale, we assessed the credibility of HTE by adapting the ICEMAN criteria for RCTs (21,90). Although these criteria were developed for assessing credibility of findings from one-at-a-time subgroup analyses of interactions, three of the five apply readily to multivariable predictive modeling. These include 1) whether or not the number of interactions tested is small (three or fewer); 2) whether interactions tested are limited to covariates for which prior evidence of possible effect modification exists; and 3) whether arbitrary or data-driven cut-points are avoided in analyzing possible treatment interactions with continuous covariates.

Each ICEMAN criterion (90) rates compliance from 1 to 4 (definitely not, probably not, probably and definitely compliant). If all criteria are scored as probably or definitely compliant, credibility of HTE is rated as "high". If at least 2 criteria are scored as definitely not compliant or if all 3 are scored as probably not compliant or worse, credibility is rated as "very low" or "low", respectively. Remaining reports are rated as either "low" or "moderate", using reviewer discretion with guidance from the ICEMAN manual considering the quality of the methods employed and whether statistical tests, when present, supported a hypothesis of HTE. High or moderate ratings were classified as credible HTE for this review.

Two ICEMAN criteria were not readily applicable to predictive modeling. The criterion that a statistical test for interaction be performed and highly significant, was not always applicable. In risk modeling, important HTE could be identified on the absolute scale even if testing for interaction on the relative scale was null or not performed. In effect modeling, multiple possible interactions were usually tested and results for specific interactions were often not reported. When present, results of statistical tests for interaction or overall HTE were considered and may have weighed in differentiating between low and moderate credibility. The fifth criterion, that authors pre-specify direction of the interaction, was not considered feasible in predictive modeling, given the potentially complex interactions of multiple covariates with each other and with treatment.

Reviewers assessed reports of credible HTE for clinical importance by determining whether observed differences in size and direction of absolute treatment effect between subgroups supported different treatment recommendations. An additional consideration was whether findings for all outcomes studied, including adverse effects of treatment, were consistent in identifying preferred treatments, or whether they conflicted.

Results

General Description. The 42 reports (35-79) included 35 peer-reviewed publications, 4 postings on pre-print archives, and 3 dissertations. Five appeared in 2020, 11 in 2021, 13 in 2022 and 13 in the first five months of 2023. Among the 42, 25 were re-analyses of single RCTs, 14 were individual patient data meta-analyses (IPDMAs) of two or more RCTs, and three were initial reports from single RCTs that included HTE analyses. Forty-one reports examined HTE for a clinical treatment and one (54) evaluated behavioral interventions to boost educational performance among university students. A total of 19 risk models and 26 effect models were reported, with three reports presenting both risk and effect models (49,73,75).

Reviewer Agreement. After excluding six reports (presenting six effect models and one risk model) used for training reviewers, initial between-reviewer disagreement rates for doubly-reviewed items in 36 reports ranged from 0 to 47%, with an overall average of 17.4% (details, Supplement Table S2). Disagreements were resolved with discussion. Items with higher levels of initial disagreement included both the credibility and clinical importance of claimed HTE, especially for risk models.

Risk Models. Risk modeling appeared with increasing frequency over time, with six appearing during 2020-21 and an additional 13 found in 2022 and the first five months of 2023. Consistency with Statement criteria (Table 1) was above 65% for all but three criteria. Slightly fewer than half of reports with positive findings included a risk model. External prediction models were employed in only eight of 19 analyses, possibly because an appropriate validated model was not available; and risk score distributions were presented separately by trial arm in only 12 reports.

Study authors claimed findings of HTE in 15 risk modeling reports (Figure 2). For six, (49,60,64,71,75,77) heterogeneity was found on the absolute but not the relative scale (i.e., benefit magnification) and for nine (36,38,43,48,50,54,58,68,73), heterogeneity was also found on the relative scale (i.e., relative treatment effects also varied across levels of baseline risk). The

four remaining reports (56,57,61,76) did not find clear overall treatment effects and none claimed HTE on either scale. Risk models generally scored highly on the three ICEMAN criteria for credibility of HTE. They always involved a single effect modifier (the baseline risk score); those finding benefit magnification had strong prior theoretical and empirical reasons (18) to expect such HTE; and all models used pre-specified rather than data-driven cut-points to define risk score subgroups.

We scored all risk models finding benefit magnification and eight of nine that found HTE on a relative scale as credible HTE. Four of these eight found that relative as well as absolute treatment effects were greater in individuals at higher risk for experiencing trial outcomes. (38,43,50,68) In the remaining four (36,48,58,73), those at greatest risk showed no evidence of benefit. In two (36,58), both relative and absolute effects of treatment were greatest for individuals in the middle of the risk distribution, and in two (48,73), only those at lower risk experienced a treatment benefit. In the single report that was not found to be credible, multiple treatment-by-risk interactions were tested across a variety of outcomes and findings were inconsistent. (54)

Effect Models. The 26 effect model analyses used diverse types of models and analytic strategies. Nine analyses used regression methods; the remainder employed various data-driven machine-learning approaches. Machine-learning approaches became more frequent over time (5/11 reports in 2020-21 vs. 12/15 reports in 2022-23).

Few effect model reports (41,44,59) restricted analyses to potential effect modifiers with strong prior evidence for effect modification (Table 3). The majority explored many candidate effect modifiers with little prior evidence. Most used recommended steps to reduce risks of over-fitting, including coefficient shrinkage methods and internal cross-validation. Six (40,41,44,59,69,79) applied effect model findings to external datasets for validation. Only four studies (52,75,78,79) used performance metrics designed for risk prediction without also reporting performance for

predicting treatment effects. In all, eight effect modeling reports (41,42,44,49,59,62,69,74) specifically assessed model performance for predicting individual treatment effects.

Authors claimed HTE in 20 effect modeling reports (Figure 2). Most failed to meet the adapted ICEMAN credibility criteria because they explored many variables with little prior evidence for effect modification. Many also reported data-driven rather than pre-specified cut points for continuous predictors. However, five of the 20 were judged to present credible HTE. These included three (41,44,59) that restricted analyses to small numbers of pre-specified effect modifiers with strong prior evidence. These, along with two other reports, (40,97) also validated model predictions of individual treatment effect in external RCT datasets. Because these validation analyses tested only effect modifiers with prior evidence and pre-specified cut points (i.e., the evidence and cut points from derivation analyses), they scored highly for credibility of HTE. Four of these used regression models; (40,41,44,59) the fifth (69) used a causal forest machine-learning algorithm. (82)

Assessment for Clinically Important HTE. Reviewers judged findings from 14 reports with credible HTE to also be clinically important (Figure 2, Table 3), including 10 of 14 risk modeling analyses and four of five effect modeling reports. Table 3 gives the decision thresholds for important subgroup differences in treatment recommendations. Reasons for concluding that credible HTE was not clinically important included lack of clarity in presentation of findings, (75) failure to identify a threshold for differing treatment choices, (64) conflicting findings across outcomes, (77) failure to add clinical value to previous risk-based selection strategies, (68) and concurrence with authors on the need for additional investigation, possibly testing additional potential effect modifiers. (44)

Discussion

Popular approaches to evidence-based medicine have encouraged reliance on average treatment effects from RCTs to support decision-making by individual patients, (91) despite appreciation of the limitations of this approach. Herein, we reviewed early efforts in applying predictive modeling within RCTs to deliver more patient-centered evidence for decision making. During the first three years, five months following publication of the PATH Statement, we identified 42 reports that cited the Statement and presented predictive modeling across a range of clinical conditions and types of interventions. Fully one third of these reports found HTE that met adapted ICEMAN criteria for credibility and the PATH definition for clinical importance, providing strong evidence that recommendations should vary among patients facing the same treatment choices.

The Statement recommended that risk modeling be conducted when RCTs report positive overall results. Risk modeling was much more likely than effect modeling to produce findings that met criteria for credible HTE because risk modeling tests only a single effect modifier, one with strong prior evidence and a theoretical rationale for effect modification. Nevertheless, fewer than half of reports from positive RCTs presented risk modeling. Many effect modeling reports had features the Statement indicated would make risk-modeling a promising place to begin (Supplement, Box A, Table S4), suggesting that a simpler approach could have been more informative.

Contrary to assumptions that relative treatments effects are constant across levels of baseline risk, the risk modeling studies reviewed here more often found that relative as well as treatment effects varied importantly. In one report, (73) persons at opposite ends of the predicted risk range experienced opposite effects of treatment. In others, relative effects were greater in or completely confined to persons at higher (38,43,50,68) or lower ends (48) of predicted risk, and in two, (36,58) maximal benefit was found for those in the mid-range. This U-shaped, or “sweet spot,” pattern (36) has also been observed elsewhere. (92)

Explanations for such variation in relative treatment effects across baseline risk are not always obvious. Risk scores may incorporate traits that are both strong predictors of study outcomes and also potent relative treatment effect modifiers, either directly or as proxies for unmeasured attributes. In a study of therapeutic-dose heparin vs. usual care pharmacologic thromboprophylaxis for patients hospitalized with COVID-19, (73) a better initial respiratory status was the most potent predictor of good clinical outcomes. When included in the risk model, respiratory status dominated the risk score. It had also proved to be a strong modifier of treatment effect in earlier subgroup analyses. Persons with low risk scores (better baseline respiratory status) were then found to be the only subgroup that benefited from heparin treatment. In three RCTs (36,48,58) where incidence of study outcomes was particularly high (range 27-61%), no benefit was observed for those in the highest stratum of predicted risk. For these extremely high-risk individuals, models may have captured attributes whose presence reflected irreversible disease or competing causes of the outcome that would make treatment futile. These observations of risk – treatment interactions and others noted elsewhere (93) demonstrate that assumptions of simple benefit magnification are not well-founded and should be tested routinely.

The number of effect modeling analyses and the increasing use of exploratory machine-learning methods over time suggest continuing enthusiasm for individualizing treatment recommendations beyond risk stratification. Several authors motivated their approaches by pointing to limitations of risk modeling. (47,65,77) Although risk scores can create patient subgroups well-matched on risk, subgroup members may be heterogeneous for the specific characteristics that contributed to their risk scores and therefore potentially heterogeneous in terms of their treatment response. (65)

Five effect models did find credible HTE, either because they adhered to Statement recommendations to include only effect modifiers with strong prior evidence or because they

took the added step of validating their HTE findings in external populations. Many additional reports provided evidence suggesting multivariable HTE that now deserves such external validation.

Findings from several effect modeling reports revealed the uncertainty that can remain after initial findings suggest HTE and reinforce the necessity of validating such findings in other populations before they are accepted as credible. Two reports (47,74) explored data from the SPRINT and Action to Control Cardiovascular Risk in Diabetes (ACCORD) trials using causal forest algorithms. One (74) found evidence of HTE, the other did not. In two reports (52,70) from a trial of dabigatran vs. warfarin for stroke prevention in atrial fibrillation, one model (52) suggested interactions of three covariates with treatment choice and significant HTE; the second report, (70) using four machine-learning algorithms applied to the same RCT data, found no evidence for HTE. Sinha et al (51) applied four widely used machine-learning algorithms to RCTs of treatments for acute respiratory distress syndrome finding inconsistent evidence for HTE, not only between algorithms but within algorithms when random initiation seeds were altered. Both authors (51,70) acknowledged the challenges from false signals of effect modification in exploratory analyses.

As the Statement suggests, additional steps will often be needed before implementing HTE findings from predictive models into clinical practice. Prediction models used in risk modeling, especially those developed internally, may not yet generalize well to clinical populations with differing risk distributions. The performance of all models for predicting treatment effects may need further validation in differing populations. Ultimately, it will be critical to demonstrate that clinical outcomes improve when treatment recommendations are personalized using predictive modeling.

There are learnings from this review for funding, conducting, and publishing clinical research. The value of external validation, especially for effect models, points to the fundamental

importance of sharing data from completed RCTs. In the absence of additional appropriate RCTs, large well-characterized observational cohorts with treatment and covariate data could also be valuable both for validation (94) as well as for developing and validating new, representative risk prediction models. In planning and funding new RCTs, the potential existence of appropriate external risk models should be considered and data collection should include baseline data necessary for classifying individual risk of study outcomes.

In recent years, guidelines for reporting clinical trials findings have recognized the importance of presenting absolute as well as relative measures of the overall treatment benefit or harm because of the greater relevance of absolute measures to clinical decision-making. (95-97) We believe the present review supports consideration of an additional editorial requirement that initial reports of RCTs or IPDMA's routinely present treatment effects in relation to baseline risk when overall results are positive.

Limitations.

There is inherent subjectivity in assessing credibility and importance of HTE. The close association of two authors (DK, JS) with production of the PATH statement should be kept in mind.

Conclusions.

The PATH statement appears to be influencing research practice. Effect modeling holds promise for predicting individualized treatment effects but the need for external validation is a constraint. Risk modeling provides a more straightforward initial approach when overall trial findings are positive and often identifies clinically important HTE.

Acknowledgments:

The authors gratefully acknowledge Harold Sox, MD, Department of Medicine and The Dartmouth Institute (emeritus), Geisel School of Medicine at Dartmouth, Hanover NH, for careful review and helpful suggestions on earlier drafts of the manuscript; Jinny G. Park, MPH, Tufts Predictive Analytics and Comparative Effectiveness Center, Tufts University School of Medicine, Boston MA, for conducting all literature database searches; and Ivan Rivera, MIS, Division of Research, Kaiser Permanente Northern CA, for retrieving reprints and supplemental materials of study citations.

Data Availability Statement:

All data produced in the present study are available upon reasonable request to the authors.

Funding:

Drs. Selby and Maas and Mr. Fireman report no funding related to work performed on this publication. Dr. Kent was funded by a National Institutes of Health (NIH)/National Center for Advancing Translational Sciences (NCATS) grant (UM1TR004398-01). Dr. Selby previously served as the Executive Director of the Patient-Centered Outcomes Research Institute (PCORI). The views and findings presented in this publication are solely the responsibility of the authors and are not presented on behalf of or as the views of PCORI.

References

1. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA. 1991; 266:93–98.

2. Rothwell PM. Can overall results of clinical trials be applied to all patients? Lancet. 1995; 345:1616–19.

3. Horwitz RI, Singer BH, Makuch RW, Viscoli CM. Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information needs of clinical inquiry and drug regulation. J Clin Epidemiol. 1996;49:395-400.

4. Feinstein AR. The Problem of Cogent Subgroups: A Clinicostatistical Tragedy. J Clin Epidemiol. 1998; 51:297-99.

5. Mant D. Can randomised trials inform clinical decisions about individual patients? Lancet. 1999;353:743-46.

6. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA. 2007; 298(10):1209–12.

7. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005; 365:176–86.

8. 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.

9. 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.

10. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117.

11. Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) National Priorities for Research and Initial Research Agenda. JAMA. 2012;307:1583-84.

12. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372:793–5

13. Patient-Centered Outcomes Research. Subtitle D of Title VI - Sec. 6301. Patient Protection and Affordable Care Act. 2010. Available from: https://www.pcori.org/sites/default/files/PCORI-Authorizing-Legislation.pdf. Accessed on Jan 22, 2024.

14. Kent DM, Paulus JK, van Klaveren D et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172:35-45.

15. 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-W25.

16. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018:363:k4245

17. Harrell F. Viewpoints on Heterogeneity of Treatment Effect and Precision Medicine. Available from: https://www.fharrell.com/post/hteview/index.html. Accessed on Jan 22, 2024.

18. 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; 1(45):2075–88.

19. 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

20. Tricco AC, Lillie E, Zarin W et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169:467-73.

21. Schandelmaier S, Briel M, Varadhan R et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020;192:E901-06.

22. Kataoka H, Mochizuki T, Ohara M et al; FEATHER Investigators. Urate-lowering therapy for CKD patients with asymptomatic hyperuricemia without proteinuria elucidated by attribute-based research in the FEATHER Study. Sci Rep. 2022;12:3784-96.

23. Albuquerque AM, Tramujas L, Sewanan LR, Brophy JM. Mortality Rates Among Hospitalized Patients With COVID-19 Infection Treated With Tocilizumab and Corticosteroids: A Bayesian Reanalysis of a Previous Meta-analysis JAMA Netw Open. 2022; 5:e220548.

24. Foy AJ, Filippone EJ, Schaefer E et al. Association Between Baseline Diastolic Blood Pressure and the Efficacy of Intensive vs Standard Blood Pressure-Lowering Therapy. JAMA Netw Open. 2021;4:e2128980.

25. Kloecker DE, Khunti K, Davies MJ et al. Microvascular Disease and Risk of Cardiovascular Events and Death From Intensive Treatment in Type 2 Diabetes: The ACCORDION Study. Mayo Clin Proc. 2021;96:1458-1469.

26. Dianti J, McNamee JJ, Slutsky AS et al. Determinants of Effect of Extracorporeal CO_2 Removal in Hypoxemic Respiratory Failure. NEJM Evid. 2023;2(5). Available from: https://evidence.nejm.org/doi/full/10.1056/EVIDoa2200295. Accessed on Jan 22, 2024.

27. Farrar J, Locke K, Clemens J et al. Widespread Pain Phenotypes Impact Treatment Efficacy Results in Randomized Clinical Trials for Interstitial Cystitis/ Bladder Pain Syndrome: A MAPP Network Study. ResearchSquare. Available from: https://www.researchsquare.com/article/rs-2441086/v1. Accessed on Jan 22, 2024.

28. Hanlon P, Butterly EW, Shah ASV et al. Treatment effect modification due to comorbidity: Individual participant data meta-analyses of 120 randomised controlled trials. PLOS Medicine. 2023; Accessed September 27, 2023: https://doi.org/10.1371/journal.pmed.1004176. Accessed on Jan 22, 2024.

29. Samuels N, van de Graaf RA, Mulder MJHL et al; HERMES Collaborators. Admission systolic blood pressure and effect of endovascular treatment in patients with ischaemic stroke: an individual patient data meta-analysis. Lancet Neurol. 2023;22:312-19.

30. Kimchi A, Aronow HU, Ong MK et al; BEAT-HF Research Group. Post-discharge Noninvasive Telemonitoring and Nurse Telephone Coaching Improve Outcomes in Heart Failure Patients With High Burden of Comorbidity. J Card Fail. 2023;29:774-83.

31. Gargiulo G, Giacoppo D, Jolly SS et al; Radial Trialists Group. Effects on Mortality and Major Bleeding of Radial Versus Femoral Artery Access for Coronary Angiography or Percutaneous Coronary Intervention: Meta-Analysis of Individual Patient Data From 7 Multicenter Randomized Clinical Trials. Circulation. 2022;146:1329-43.

32. Klitgaard TL, Schjørring OL, Lange T et al. Lower versus higher oxygenation targets in critically ill patients with severe hypoxaemia: secondary Bayesian analysis to explore heterogeneous treatment effects in the Handling Oxygenation Targets in the Intensive Care Unit (HOT-ICU) trial. Br J Anaesth. 2022;128:55-64.

33. Wijn SRW, Hannink G, Osteras H et al. Arthroscopic partial meniscectomy vs non-surgical or sham treatment in patients with MRI-confirmed degenerative meniscus tears: a systematic review and meta-analysis with individual participant data from 605 randomised patients. Osteoarthritis Cartilage. 2023;31:557-66.

34. Inoue K, Hsu W, Arah OA et al. Generalizability and Transportability of the National Lung Screening Trial Data: Extending Trial Results to Different Populations. Cancer Epidemiol Biomarkers Prev. 2021;30:2227-34.

35. Redelmeier D, Tibshirani RJ. An approach to explore for a sweet spot in randomized trials. J Clinical Epidemiol. 2020;120:59-66.

36. Redelmeier DA, Thiruchelvam D, Tibshirani RJ. Testing for a Sweet Spot in Randomized Trials. Med Decis Making. 2022;42:208-16.

37. Chalkou K, Steyerberg E, Egger M et al. A two□stage prediction model for heterogeneous effects of treatments. Statistics in Med. 2021;40:4362–75.

38. Chalkou K, Hamza T, Benkert P et al. Combining randomized and non-randomized data to predict heterogeneous effects of competing treatments. Stat ArXiv. Available from: https://doi.org/10.48550/arXiv.2302.14766. Accessed on Jan 22, 2024.

39. Troxel AB, Petkova E, Goldfeld K et al. Association of Convalescent Plasma Treatment With Clinical Status in Patients Hospitalized With COVID-19: A Meta-analysis. JAMA Netw Open. 2022;5:e2147331.

40. Park H, Tarpey T, Liu M et al. Development and Validation of a Treatment Benefit Index to Identify Hospitalized Patients With COVID-19 Who May Benefit From Convalescent Plasma. JAMA Network Open. 2022; 2022;5(1):e2147375.

41. Takahashi K, Serruys PW, Fuster V et al on behalf of the SYNTAXES, FREEDOM, BEST, and PRECOMBAT trial investigators. Redevelopment and validation of the SYNTAX score II to

individualise decision making between percutaneous and surgical revascularisation in patients with complex coronary artery disease: secondary analysis of the multicentre randomised controlled SYNTAXES trial with external cohort validation. Lancet. 2020;396: 1400-12.

42. Nguyen TL, Collins GS, Landais G. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. Journal of Clinical Epidemiology. 2020;125:47e56.

43. Kumar V, Shaw JR, Key NS et al. D-Dimer Enhances Risk-Targeted Thromboprophylaxis in Ambulatory Patients with Cancer. Oncologist. 2020;25:1075-83.

44. Dennis JM. Precision Medicine in Type 2 Diabetes: Using Individualized Prediction Models to Optimize Selection of Treatment. Diabetes. 2020;69:2075-85.

45. Rudolph KE, Díaz I, Luo SX et al. Optimizing opioid use disorder treatment with naltrexone or buprenorphine. Drug Alcohol Depend. 2021;228:1090-31.

46. Fazzari MJ, Kim MY. Subgroup discovery in non-inferiority trials. Stat Med. 2021;40:5174-75.

47. Yadlowsky S, Fleming S, Shah N et al. Evaluating Treatment Prioritization Rules via Rank Weighted Average Treatment Effects. arXiv:2111.07966v1 [stat.ME] 15 Nov 2021 Available from: https://arxiv.org/abs/2111.07966. Accessed on Jan 22, 2024.

48. Rysavy MA, Li L, Tyson JE et al; Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network. Should Vitamin A Injections to Prevent Bronchopulmonary Dysplasia or Death Be Reserved for High-Risk Infants? Re-analysis of the National Institute of Child Health and Human Development Neonatal Research Network Randomized Trial. .J Pediatr. 2021;236:78-85.

49. Bress AP, Greene T, Derington CG et al; SPRINT Research Group. Adverse Patient Selection for Intensive Blood Pressure Management Based on Benefit and Events. .J Am Coll Cardiol. 2021;77:1977-90.

50. Kent DM, Saver JL, Kasner S et al. Heterogeneity of Treatment Effects in an Analysis of Pooled Individual Patient Data From Randomized Trials of Device Closure of Patent Foramen Ovale After Stroke. .JAMA. 2021;326:2277-86.

51. Sinha P, Spicer A, Delucchi KL et al. Comparison of machine learning clustering algorithms for detecting heterogeneity of treatment effect in acute respiratory distress syndrome: A secondary analysis of three randomised controlled trials. EBioMedicine. 2021;74:103697. Available from: https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00491-6/fulltext. Accessed on Jan 22, 2024.

52. Reinhardt SW, Desai NR, Tang Y et al. Personalizing the decision of dabigatran versus warfarin in atrial fibrillation: A secondary analysis of the Randomized Evaluation of Long-term anticoagulation therapY (RE-LY) trial. PLoS One. 2021;16:e0256338.

53. Kessler RC, Furukawa TA, Kato T et al. An individualized treatment rule to optimize probability of remission by continuation, switching, or combining antidepressant medications after failing a first-line antidepressant in a two-stage randomized trial. Psych Med. 2021;8:1-10.

54. Brade R. Behavioral Interventions and Students' Success at University: Evidence from Randomized Field Experiments. Dissertation, University of Gottingen, 2021. Available from: https://ediss.uni-goettingen.de/bitstream/handle/21.11130/00-1735-0000-0008-59F8-D/Brade_Dissertation.pdf?sequence=1. Accessed on Jan 22, 2024.

55. Edward JA, Josey K, Bahn G et al.Heterogeneous treatment effects of intensive glycemic control on major adverse cardiovascular events in the ACCORD and VADT trials: a machine-learning analysis. Cardiovas Diabetol. 2022;21:58.

56. Stefano LD, Ogburn EL, Ram M et al; Pandemic Response COVID-19 Research. Hydroxychloroquine/ Chloroquine for the Treatment of Hospitalized Patients with COVID-19: An Individual Participant Data Meta-Analysis. PLoS ONE; 2022; 17(9): e0273526. Available from: https://doi.org/10.1371/journal.pone.0273526. Accessed on Jan 22, 2024.

57. Granholm A, Munch MW, Myatra SN et al. Dexamethasone 12 mg versus 6 mg for patients with COVID-19 and severe hypoxaemia: a pre-planned, secondary Bayesian analysis of the COVID STEROID 2 trial. Intensive Care Med. 2022;48:45-55.

58. Taylor SP, Murphy S, Rios A et al. Effect of a multicomponent sepsis transition and recovery program on mortality and readmissions after sepsis: The Improving Morbidity During Post-Acute Care Transitions for Sepsis Randomized Clinical Trial. Critical Care; 2022;50:469-479.

59. Dennis JM, Young KG, McGovern AG et al; on behalf of the MASTERMIND Consortium. Development of a treatment selection algorithm for SGLT2 and DPP-4 inhibitor therapies in people with type 2 diabetes: a retrospective cohort study. Lancet Digit Health; 2022;4:e873-83.

60. Gencer B, Eisen A, Berger D et al. Edoxaban versus Warfarin in high-risk patients with atrial fibrillation: A comprehensive analysis of high-risk subgroups. Am Heart J. 2022;247:24-32.

61. Pinho-Gomes AC. Management of blood pressure in atrial fibrillation, heart failure and multimorbidity. Dissertation, Oxford University 2020. Available from: https://ora.ox.ac.uk/objects/uuid:fcbe8b1d-4846-4499-95ef-b7ba3b5ef9a3. Accessed on Jan 22, 2024

62. Chen X, Harhay MO, Tong G, Li F. A Bayesian Machine-Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial. arXiv:2204.06657v1 [stat.AP] 13 Apr 2022. Available from: https://arxiv.org/pdf/2204.06657.pdf. Accessed on Jan 22, 2024.

63. Wolf JM , Koopmeiners JS, Vock DM. A permutation procedure to detect heterogeneous treatment effects in randomized clinical trials while controlling the type I error rate. Clin Trials. 2022;19:512-21.

64. van Kruijsdijk RCM, Vernooij RWM, Bots ML et al; HDF Pooling Project investigators. Personalizing treatment in end-stage kidney disease: deciding between haemodiafiltration and haemodialysis based on individualized treatment effect prediction Clin Kidney J. 2022;15:1924-31.

65. Nguyen T-L, Trompet S, Broderson JB et al. The potential benefit of statin prescription based on prediction of treatment responsiveness in older individuals: An application to the PROSPER randomised controlled trial. Eur J Prev Cardiol. 2023 Dec 12:zwad383. doi: 10.1093/eurjpc/zwad383. Online ahead of print.

66. Sadique Z , Grieve R, Diaz-Ordaz K et al. A Machine-Learning Approach for Estimating Subgroup- and Individual-Level Treatment Effects: An Illustration Using The 65 Trial. Med Decis Making. 2022;42:923-36.

67. Rudolph KE, Williams NT, Díaz I et al. Optimally Choosing Medication Type for Patients With Opioid Use Disorder. Am J Epidemiol. 2023;192:748-56.

68. Mell LK, Pugh SL, Jones CU et al. Effects of Androgen Deprivation Therapy on Prostate Cancer Outcomes According to Competing Event Risk: Secondary Analysis of a Phase 3 Randomised Trial. Eur Urol. 2023;S0302-2838(23)00056-8.

69. Seitz KP, Spicer AB, Casey JD et al. Individualized Treatment Effects of Bougie versus Stylet for Tracheal Intubation in Critical Illness. Am J Respir Crit Care Med. 2023;207:1602-11.

70. Xu Y, Bechler K, Callahan A, Shah H. Principled estimation and evaluation of treatment effect heterogeneity: A case study application to dabigatran for patients with atrial fibrillation. Journal of Biomedical Informatics. 2023;143:104420.

71. Trinks-Roerdink EM, Geersing GJ, Van den Dries CJ et al. Integrated care in patients with atrial fibrillation – a predictive heterogeneous treatment effect analysis of the ALL-IN Trial. In: Trinks-Roerdink EM. Balancing risks in thromboembolic disease. (PhD Dissertation). Available from: https://dspace.library.uu.nl/bitstream/handle/1874/428070/phdthesis-withcover-emtrinksroerdink%20-%206450fdb962978.pdf?sequence=1#page=57. Accessed on Jan 22, 2024.

72. Colloca L, Dworkin RH, Farrar JT et al. Predicting Treatment Responses in Patients With Osteoarthritis: Results From Two Phase III Tanezumab Randomized Clinical Trials. Clin Pharmacol Ther. 2023;113:878-86.

73. Goligher EC, Lawler PR, Jensen TP et al. REMAP-CAP, ATTACC, and ACTIV-4a Investigators. Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19**.** JAMA. 2023;329:1066-77.

74. Inoue K, Athey S, Tsugawa Y. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. Int J Epidemiol. 2023;52:1243-56.

75. Ghazi L, Shen J, Ying J et al. Identifying Patients for Intensive Blood Pressure Treatment Based on Cognitive Benefit: A Secondary Analysis of the SPRINT Randomized Clinical Trial. JAMA Netw Open. 2023; 6:e2314443.

76. Gentle SJ, Rysavy MA, Li L et al. Heterogeneity of Treatment Effects of Hydrocortisone by Risk of Bronchopulmonary Dysplasia or Death Among Extremely Preterm Infants in the National Institute of Child Health and Human Development Neonatal Research Network Trial: A

Secondary Analysis of a Randomized Clinical Trial. JAMA Netw Open. 2023: 2023;6:e2315315.

77. Charu V, Liang JW, Chertow GM et al. Heterogeneous treatment effects of intensive glycemic control on kidney microvascular outcomes and mortality in ACCORD. J Am Soc Nephrol. 2023 Dec 11. doi: 10.1681/ASN.0000000000000272. Online ahead of print.

78. Zarski A-C, Harrer M, Kuper P et al. Predicting Individualized Effects of Internet-Based Treatment for Genito-Pelvic Pain/Penetration Disorder: Development and Internal Validation of a Multivariable Decision Tree Model. MedRxiv 2023; Available from: https://arxiv.org/abs/2303.08732. Accessed on Jan 22, 2024.

79. Harrer M, Ebert DD, Kuper P et al. Predicting heterogeneous treatment effects of an Internet-based depression intervention for patients with chronic back pain: Secondary analysis of two randomized controlled trials. Internet Interv. 2023 Jun 7:33:100634.

80. Su X, Tsai CL,Wang H et al. Subgroup analysis via recursive partitioning. J Mach Learn Res. 2009;10:141–58.

81. Loh WY, He X and Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med. 2015;34(11):1818–33.

82. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences. 2016;113:7353-60.

83. Luedtke AR, van der Laan MJ. Super-Learning of an Optimal Dynamic Treatment Rule**.** Int J Biostat. 2016;12:305-32.

84. Tibshirani R. The Lasso method for variable selection in the Cox model. Stat Med. 1997;16:385-95.

85. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970:12:55-67.

86. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20.

87. van Klaveren D, Steyerberg EW, Serruys PW, Kent DM. The proposed 'concordance-statistic for benefit' provided a useful metric when modeling heterogeneous treatment effects. J Clin Epidemiol. 2018:94:59-68.

88. Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment. arXiv:2209.06101v1 [stat.ME] 13 Sep 2022. Available from: https://arxiv.org/abs/2209.06101. Accessed on Jan 22, 2024.

89. Radcliffe NJ. Using control groups to target on predicted lift: Building and assessing uplift models. Direct Marketing Analytics Journal. 2007;1(3):14-21.

90. Instrument to assess the credibility of effect modification analyses (ICEMAN) in a randomized controlled trial. Available from: https://www.iceman.help/. Accessed on Jan 22, 2024.

91. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? Evidence-Based Medicine Working Group. *JAMA* 1994;271:59-63.

92. Marafino BJ, Schuler A, Liu VX et al. Predicting preventable hospital readmissions with causal machine learning. Health Serv Res. 2020;55:993–1002.

93. Rekkas A, Rijnbeek PR, Kent DM et al. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk□based approaches. BMC Med Res Methodol. 2023;**23**, 74. Available from: https://doi.org/10.1186/s12874-023-01889-6. Accessed on Jan 22, 2024.

94. Segal JB, Varadhan R, Groenwold RHH, et al. Assessing Heterogeneity of Treatment Effect in Real-World Data**.** Ann Intern Med. 2023;176:536-544.

95. Schulz KF, Altman DG, Moher D for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Ann Intern Med. 2010;152:726-32.

96. The New England Journal of Medicine. Author Center. Statistical Reporting Guidelines for New Manuscripts Author Center. Available from: https://www.nejm.org/author-center/new-manuscripts. Accessed on Jan 22, 2024.

97. The British Medical Journal. BMJ Guidance for Authors. Available from: https://www.bmj.com/sites/default/files/attachments/resources/2018/05/BMJ-InstructionsForAuthors-2018.pdf. Accessed on Jan 22, 2024.

Table 1. Consistency with PATH Statement criteria for risk modeling (19 risk models presented)

Recommendation	Number Adherent / Number Eligible*
1. Conduct a risk model analysis if the trial has positive overall results	15/32 (47%)[†]
2. Apply an externally developed risk model to stratify the trial population, if available	8/19 (42%)
3. If developing an internal model, avoid using control group only	8/11 (73%)[‡]
4. Pre-specify plan for applying/developing the model	14/19 (74%)
5. Report metrics for model performance on trial population	15/19 (70%)
6. Report distribution of predicted risk/risk score for each arm of trial	12/19 63%)
7. Report treatment effects in both relative and absolute terms across risk strata	16/19 (84%)
8. If there are important treatment-related harms (n=9), report these by risk stratum	6/9 (67%)[§]

- Several recommendations apply only to subset of reports, as indicated below.

[†] Denominator includes all reports from RCTs with positive overall findings, whether a risk model was presented or not. Numerator excludes 3 risk models presented in trials with negative overall findings.

[‡] Denominator excludes the eight reports that assigned risk based on external prediction models

[§] Denominator includes only those RCTs in which at least one treatment carried risks of important harms.

Table 2. Consistency with PATH Statement criteria for effect modeling (26 effect models presented)

Recommendation	Number Adherent / Number Eligible
1. Incorporate highly credible effect modifiers into prediction models using multiplicative interaction terms	3/26 (12%)
2. Avoid regression models that do not take into account model complexity	
a. Use shrinkage methods	21/26 (81%)
b. Use internal validation	22/26 (85%)
c. Use external validation	6/26 (23%)
3. Avoid evaluations of treatment effect model performance that use only conventional metrics for prediction of risk	22/26 (85%)
4. Report model performance in terms of ability to predict treatment effect	8/26 (31%)

Table 3. Studies found to have clinically important heterogeneity of treatment effects (HTE)

RISK MODELS							
Ref	**Clinical Condition**	**Outcome(s)**	**Randomized Intervention (s)**	**Overall RCT Findings**	**Methods**	**Type of HTE Identified**	**Clinically Important Subgroup Differences**
Kent et al[50].	Patent foramen ovale (PFO) - associated stroke	Recurrent stroke	Percutaneous PFO-closure vs medical therapy	Strong benefit in favor of PFO-closure (aHR=0.41)	IPD meta-analysis of 6 trials; external risk model	Relative as well as absolute effect variation, p=0.02 for multiplicative interaction	Identified a subgroup comprising 15% of trial population unlikely to have had a PFO-related stroke, who received no benefit from PFO-closure and were at higher risk of procedure-related complications.
Kumar et al.[43]	Ambulatory patients with cancer	Venous thrombo-embolism (VTE)	Apixaban vs placebo	Strong benefit in favor of apixaban (aHR=0.49)	Single RCT re-analysis; external risk model	Probable relative as well as absolute effect variation; test for multiplicative interaction not reported	Patients with a baseline risk for VTE 8%, comprising 67% of the trial population, derived no benefit from apixaban, but experienced an excess of overall bleeding events on treatment with apixaban
Bress et al.[49]	Systolic hypertension and increased cardiovascular risk	Cardiovascular disease events; all-cause mortality	Intensive vs standard systolic blood pressure control	Strong benefit in favor of intensive systolic control (HR=0.75 and 0.73, respectively	Single RCT re-analysis; internal risk model	Absolute but not relative effect variation (benefit magnification)	Patients in highest risk quartile clearly benefit from intensive systolic BP control with acceptable adverse event rates; those in lowest risk quartile can expect little benefit from intensive control, despite increased costs, burden and adverse effects.
Redelmeier et al.[35,36]	Congestive Heart Failure	All-cause mortality	Implantable defibrillator (ICD) vs medical management	Strong benefit in favor of ICD (OR=0.69)	Single RCT re-analysis; internal risk model	Relative as well as absolute effect variation, p<0.001 for multiplicative interaction	Mortality Benefit of ICD implantation largely confined to patients in midrange (3rd and 4th quintiles) of predicted risk, i.e., a "sweet spot"
Taylor et al.[58]	Hospitalized Patients with Sepsis	30-day mortality and readmission	Nurse-navigator led Sepsis Transition and Recovery (STAR) intervention vs usual care	Moderate benefit in favor of the intervention (aOR=0.80)	Single RCT re-analysis; external risk model	Relative as well as absolute effect variation; 95% confidence intervals for quartile-specific OR's do not overlap	Mortality and re-admission benefit of the intervention confined to patients in the middle two quartiles of predicted risk lowest (i.e., "sweet spot"). Intervention is associated with greater costs.

Ref	Clinical Condition	Outcome(s)	Randomized Interventions	Overall RCT Findings	Methods	Type of HTE Identified	Clinically Important Subgroup Differences
Chalkou et al.[37,38]	Multiple Sclerosis	Relapse of MS	3 immunologic therapies (DF: Dimethyl fumarate, GA: Glatiramer acetate, N: Natalizumab) vs placebo	Strong benefit in favor of Natalizumab (OR $_{\text{vs placebo}}$: DF: 043 GA: 0.53 N: 0.28)	IPDMA – 3 trials, using network meta-analysis; external risk model	Probable relative as well as absolute effect variation; test for multiplicative interaction not reported	Patients with baseline risk < 30%, comprising 25% of trial population, had negligible added benefit of natalizumab vs. dimethyl fumarate. Natalizumab is associated with rare but possibly fatal complication, progressive multifocal leukoencephalopathy (PML).
Rysavy et al.[48]	Extreme Prematurity	Broncho-pulmonary dysplasia	Vitamin A vs sham injection	Weak benefit in favor of vitamin A (RR=0.89)	Single RCT re-analysis; external risk model	Relative as well as absolute effect variation, p=0.03 for multiplicative interaction	Benefits of Vitamin A therapy largely confined to lowest 50% of predicted risk; no evidence for benefit among infants in highest quarter of predicted risk.
Gencer et al.[60]	Atrial Fibrillation	Net composite of stroke/ systemic embolism, major bleed, all-cause death	Lower (LDER) vs higher (HDER) dose regimens of edoxaban vs warfarin	Moderate benefit for either dose of edoxaban vs warfarin (HR=0.83, 0.89 for LDER,HDER)	Single RCT re-analysis; risk stratification based on count of number of high-risk features	Absolute effect variation (p=0.001) but not relative effect variation (p=0.065 for multiplicative interaction) (benefit magnification)	Absolute benefits of either dose of edoxaban vs warfarin increase for both stroke/embolism and major bleeding endpoints as number of risk factors increases. Negligible benefits in those with 0-1 risk factors may not justify switching in those well managed on warfarin.
Trinks-Roerdink et al [71]	Atrial Fibrillation	All-cause mortality	Integrated atrial fibrillation care vs. usual care	Strong benefit of integrated care (aHR=0.55)	Single RCT re-analysis, external risk model	Absolute but no relative effect variation (p=0.93 for multiplicative interaction) (benefit magnification)	Large benefit for persons in highest 25% of risk distribution; minimal benefit despite increased costs in lower 75%.
Goligher et al [73]	Hospitalized COVID-19 infection	Organ-support free days; hospital survival	Therapeutic-dose heparin vs usual pharmacologic thromboprophylaixis	No benefit in overall population (OR for benefit 1.05)	IPDMA – 3 trials; re-analysis; subgroup analyses, internal risk model and effect model)	Relative as well as absolute effect variation, interaction not tested for risk model; p=0.05 for non-homogeneity of risk differences by decile of predicted risk differences in effect model	Clear benefit for 60% of persons in lowest deciles of risk score (not requiring organ support at baseline); apparent harm for those needing intensive care at baseline (highest 3 deciles of risk score); similar HTE suggested in subgroup analyses and the effect model
EFFECT MODELS							
Ref	**Clinical**	**Outcome(s)**	**Randomized**	**Overall RCT**	**Method**	**Type of HTE**	**Clinically Important Subgroup**

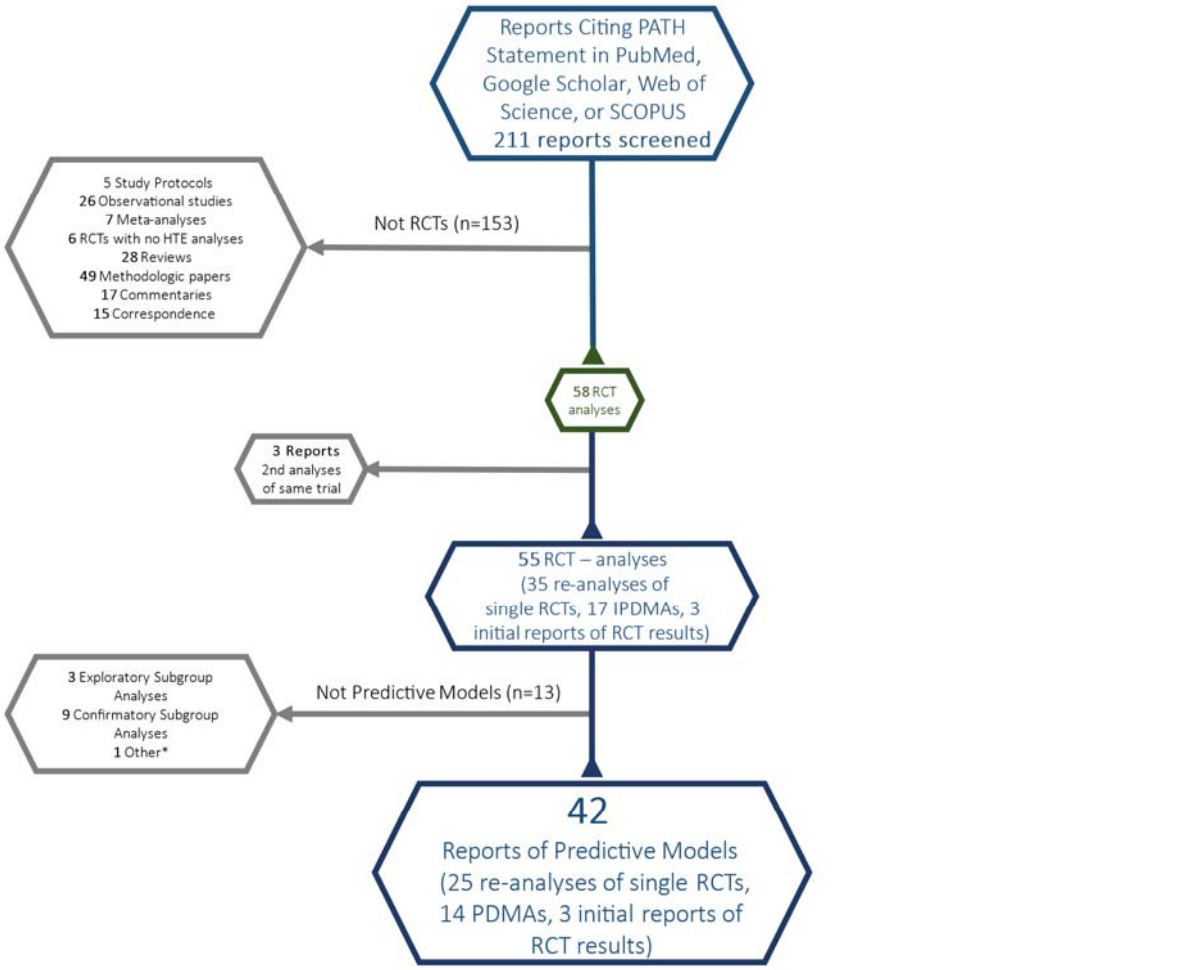
	Condition		Intervention (s)	Findings		Identified	Differences
Dennis et al.[59]	Type 2 diabetes	Hb A1c at 6 mos post-initiation	Initiation of SGLT-2 inhibitors vs DPP-4 inhibitor as add-on therapy	Strong benefit in favor of SGLT-2 inhibitor	Model built in large obs. cohort; linear regression; external validation in multiple RCTs.	Clear differences in absolute treatment effects; relative effect differences not contrasted statistically	Predictive model identifies a large subgroup, comprising 60% of trial participants, who clearly do better if SGLT-2 inhibitor is initiated (vs DPP-4 inhibitor); a small group (5%) appear to do better if a DPP-4 inhibitor is initiated.
Takahashi et al.[41]	De-novo three-vessel and left main coronary artery disease	5-year major adverse cardiac events (MACE) and 10-year all cause death	Coronary artery bypass graft (CABG) vs percutaneous coronary intervention (PCI)	Strong benefit in favor of CABG ($HR_{10\text{-year death}}$=0.84; $HR_{5\text{-year adverse events}}$=0.78)	Single RCT-reanalysis; Cox proportional hazazds regression; external validation in IPDMA of 3 trials	Multiplicative interactions of treatment with two effect model covariates (p=0.03 and 0.04)	Model identifies half of patients who clearly benefit from CABG (vs PCI); and half in whom there is no expectation of greater benefit with the more invasive procedure for either 5-year outcome (MACE) or 10-year all cause death.
Seitz et al.[69]	Respiratory Distress	Successful intubation on the first attempt	Bougie vs Stylet	Null (non-significant 6.8 percent point difference in favor of stylet)	Single RCT re-analysis; causal forest; external validation in non-random subset	Absolute and relative effect differences (p=0.02 for multiplicative interaction of treatment with predicted treatment effect)..	Model identifies a quarter of patients in the validation cohort who clearly benefit from use of the stylet vs bougie; a quarter of patients with a possible marginally better result when bougie is used; with little difference in the remainder.
Park et al[40]	Hospitalized Covid Infection without mechanical ventilation at randomization	Ordinal COVID-19 clinical status scale	Convalescent Plasma (CCP) vs control	Null (no overall association between CCP and patient outcomes)	IPDMA of 8 trials; proportional odds model; external validation in multiple datasets.	Relative treatment effect differences demonstrated with non-overlapping 95% confidence intervals	Patients were successfully categorized by the model into 3 roughly equal-sized subgroups, one with high benefit from CCP, one with modest benefit from CCP, and the third with modest harm from CCP.

FIGURE LEGENDS

Figure 1. Flow Diagram for identification and screening of all reports citing the PATH Statement and for exclusion of reports not meeting study criteria for presenting a predictive model of individual treatment effects from RCT data. Abbreviations: RCT: randomized controlled trial; IPDMA: independent patient data meta-analysis.

Figure 2. Adjudicated results of review of all eligible reports for type of predictive modeling (risk or effect), for claims by authors of heterogeneity of treatment effects (HTE), for credibility of HTE (using adapted ICEMAN criteria), and for clinical importance of HTE found to be credible.

Figure 1



*Estimated impact on various populations of CTscreening for lung cancer applying inverseprobabilityweights to stratum-specific findings from large RCT

Figure 2