Potential clinical impact of predictive modeling of heterogeneous treatment effects: scoping review of the impact of the PATH Statement

Running Title: Predictive Modeling of Individualized Treatment Effects in RCTs

Joe V Selby, MD MPH [*]

Carolien C H M Maas, MSc [†,‡]

Bruce H Fireman, MA [§]

David M Kent, MD MS [†]

* 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
510 604-2778

Word Count: 3512

ABSTRACT (310 words)

Background: The PATH Statement (2020) proposed predictive modeling for examining heterogeneity in treatment effects (HTE) in randomized clinical trials (RCTs). It distinguished risk modeling, which develops a multivariable model predicting individual baseline risk of study outcomes and examines treatment effects across risk strata, from effect modeling, which directly estimates individual treatment effects from models that include treatment, multiple patient characteristics and interactions of treatment with selected characteristics.

Purpose: To identify, describe and evaluate findings from reports that cite the Statement and present predictive modeling of HTE in RCTs.

Data Extraction We identified reports using PubMed, Google Scholar, Web of Science, SCOPUS through July 5, 2024. Using double review with adjudication, we assessed consistency with Statement recommendations, credibility of HTE findings (applying criteria adapted from the Instrument to assess Credibility of Effect Modification Analyses (ICEMAN)), and clinical importance of credible findings.

Results: We identified 65 reports (presenting 31 risk models, 41 effect models). Contrary to Statement recommendations, only 25 of 48 studies with positive overall findings included a risk model; most effect models included multiple predictors with little prior evidence for HTE. Claims of HTE were noted in 23 risk modeling and 31 effect modeling reports, but risk modeling met credibility criteria more frequently (87 vs 32 percent). For effect models, external validation of HTE findings was critical in establishing credibility. Credible HTE from either approach was usually judged clinically important (24 of 30). In 19 reports from trials suggesting overall treatment benefits,

modeling identified subgroups of 5-67% of patients predicted to experience no benefit or net treatment harm. In five that found no overall benefit, subgroups of 25-60% of patients were nevertheless predicted to benefit.

Conclusions: Multivariable predictive modeling identified credible, clinically important HTE in one third of 65 reports. Risk modeling found credible HTE more frequently; effect modeling analyses were usually exploratory, but external validation served to increase credibility.

INTRODUCTION

Overall, or average, treatment effects from randomized clinical trials (RCTs) provide limited information to patients making personal treatment decisions. (1-6) Even in strongly positive RCTs, some patients do not benefit from the favored treatment and yet may experience adverse effects. Patients and clinicians would benefit greatly if more individualized evidence could be generated and reported from RCTs.

Most publications of RCT results continue to limit examination for possible heterogeneity of treatment effects (HTE) to one-at-a-time comparisons between numerous patient subgroups, e.g. men vs. women, persons with vs. without diabetes, even though guidelines for identifying HTE in RCTs (1,7-11) have long emphasized that such analyses are at great risk for both false positive and false negative findings. An additional limitation of this approach is that individuals simultaneously belong to multiple subgroups that may vary in whether or how they appear to benefit. Thus, guidelines consistently recommend limiting the number of subgroups studied to those with prior evidence or strong biologic or clinical rationale for HTE and using caution in interpreting or applying findings to clinical practice.

The emergence of precision medicine (12) and patient-centered outcomes research (13) heightens interest in identifying important HTE. In 2020, an expert panel funded by the Patient-Centered Outcomes Research Institute (PCORI) published The Predictive Approaches to Treatment Heterogeneity (PATH) Statement (14,15), which described predictive modeling approaches that incorporate multiple patient attributes simultaneously to identify HTE and predict individualized treatment effects. The

Statement pointed out that HTE, sometimes referred to as treatment effect modification and tested as a statistical interaction, should be sought on both the absolute scale (e.g., as risk *differences)* and on the relative scale (e.g., as *ratios* of risks*,* odds or hazards) and emphasized that heterogeneity in absolute treatment effects matters more to individual patients and clinicians making treatment decisions.

The Statement distinguished two approaches to predictive modeling. “Risk modeling” incorporates multiple baseline patient characteristics into a model predicting risk for the RCT’s outcome (usually the primary outcome). In a second step, both absolute and relative treatment effects are examined across pre-specified strata (e.g., quarters) of predicted risk. (16) In the second approach, “effect modeling”, a model is developed within the RCT data to directly estimate individual treatment effects by including treatment, multiple covariates and interactions of treatment with one or more covariates. Both regression methods and more flexible, non-parametric, data-driven machine-learning algorithms (e.g., 17-20) have been used in effect modeling.

The Statement recommended risk modeling whenever an RCT demonstrates an overall treatment effect. Risk of study outcomes varies substantially in most RCT populations when participants are stratified using a multivariable model. Assuming that the relative treatment effect is homogeneous in the population, the absolute benefit would be expected to increase as predicted risk increases, (21-22) This mathematical relationship has been called “risk magnification.” (23) and is often referenced in the evidence-based medicine literature (10,24,25) and is implicit in clinical guidelines that reserve treatments with high costs or potential adverse effects for those with higher baseline risk (26,27) The Statement encouraged use of validated external models for

predicting risk if available, but indicated that in their absence, models can be developed within the RCT population, using baseline covariates and observed study outcomes and including both study arms.

Although effect modeling theoretically permits a more robust examination of possible HTE, the Statement emphasized the vulnerability of these approaches to overfitting, (28) and recommended that effect modeling be used only when there are a small number of previously established effect modifiers. It also encouraged use of methods to reduce risks of "over-fitting" of data as well as validation of effect model findings in external datasets when possible.

We conducted a scoping review (29) to assess the impact of the PATH Statement in terms of the frequency with which predictive modeling analyses of RCT data have appeared and cited the PATH Statement since its publication, the consistency of analyses with Statement consensus criteria (Supplement Table 2), and the credibility and clinical importance of claimed HTE. We applied criteria adapted from the Instrument to assess Credibility of Effect Modification Analyses (ICEMAN) (11) to assess credibility, and when HTE was found to be credible, we assessed "clinical importance," using the Statement definition of "variation across patients in the absolute 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 (Supplement

Table 1A), we identified reports that appeared between January 7, 2020 and July 24, 2024, cited the Statement and presented multivariable predictive modeling in RCT data to identify HTE. We included non-peer-reviewed reports from pre-print archives and dissertations on institutional websites.

Among 312 citations identified (Figure 1), 83 (30-112) involved analyses of data from RCTs. Sixteen (30-45) were excluded because they did not examine HTE using multivariable predictive models. Supplement Table 1B presents reasons for exclusion. Three analyses were represented by two reports each (46-51) and one report (110) presented predictive models from two distinct trials in different clinical areas, leaving 65 reports analyzing data from 162 RCTs.

Review of Predictive Model Reports. Variables collected and coding instructions are presented in Supplement Tables 3 and 4, respectively. All details of analytic strategies and findings of HTE were doubly reviewed by the first author and one co-author. A “learning” set of six reports was reviewed and discussed by all co-authors. Thereafter, co-reviewers discussed and resolved initial disagreements.

Review classified each report as risk modeling, effect modeling or both and further classified effect models into those based primarily on regression methods (e.g., ordinary least squares, logistic, proportional hazards, Bayesian regression methods) and those using more flexible data-driven machine-learning algorithms. For both risk and effect modeling analyses, we noted whether authors reported having found HTE on either absolute or relative scales and whether results of statistical testing for HTE were presented. We included statistical tests for treatment-covariate interactions from

regression models, direct contrasts of treatment effects across subgroups, reporting of confidence intervals for subgroup treatment effect estimates, and overall tests of the null hypothesis of homogeneity of treatment effect in machine-learning algorithms.

We determined whether the performance of final models for predicting individual or subgroup treatment effects was validated in datasets external to the derivation population, including validations conducted in entirely distinct RCTs, those conducted in pre-specified, non-random subsets of the original RCT population (e.g., subsets selected on bases of geography (trial sites) or time of enrollment), and those conducted in large observational cohorts.

Assessment of Credible and Clinically Important HTE. To assess credibility of claimed HTE, on either absolute or relative scales, we adapted four of the five ICEMAN criteria for RCTs (11,113). Detailed description and scoring guidance for each criterion and the overall credibility score are detailed in Supplement Tables 7A and 7B. Although ICEMAN criteria were originally developed for evaluating treatment effect modification by single covariates, four apply readily to predictive modeling with multiple covariates. These include 1) Did the authors test only a small number of interactions; 2) Was possible effect modification by each covariate supported by prior evidence; 3) If the covariate is a continuous variable, were arbitrary, data-driven cut points avoided; and 4) Does a statistical test for interaction suggest that chance is an unlikely explanation of the apparent HTE? The fifth criterion, whether the direction of interaction was hypothesized in advance, is not applicable to predictive modeling, given that multiple covariates and potentially complex interactions are evaluated simultaneously. No single criterion, including that of statistical testing, is treated as either sufficient or necessary

for establishing overall credibility. Overall credibility scores range from 1 to 4 (very low, low, moderate, or high credibility).

Because risk models involve a single effect modifier (the baseline risk score) with strong prior theoretical and empirical support (21) for HTE, at least on the absolute scale; and because risk modeling either reports pre-specified risk score cut-points or treats risk as a continuous variable, risk models can be expected to score well when ICEMAN criteria are applied. Most effect models tested multiple potential treatment-covariate interactions, often with little prior evidence and therefore tended to score poorly with application of adapted ICEMAN criteria. However, we gave considerable weight to external validation of effect model performance in another population. External validation essentially tests for HTE across a single vector, the “effect score” or predicted individual treatment effect, much as risk modeling tests for HTE across the risk score. (114) Overall credibility usually rose to “moderate“ if models performed well in external validation, even if credibility of derivation analyses would have been scored as very low.

We classified all reports scored as at least moderate overall credibility as “credible” and assessed findings for clinical importance. Per the PATH Statement, clinical importance is based exclusively on the size and direction of observed differences in absolute treatment effects between subgroups and whether these differences appear sufficient to support differing treatment recommendations. An additional consideration was whether findings for all outcomes studied, including adverse effects of treatment, were consistent in supporting the same treatment choice.

Results

General Description. Predictive models of HTE appeared with increasing frequency each year following publication of the PATH Statement (Table 1). Among the 65 reports (46-112), we identified 31 risk modeling and 41 effect modeling analyses. Seven reports (60,84,86,91,105,109,111) presented both risk and effect modeling analyses. Most effect modeling reports examined large numbers of potential interactions, and the majority employed data-driven, non-parametric analytic methods.

Reviewer Agreement. Excluding six reports (presenting six effect models and one rik model) used for training reviewers, initial between-reviewer disagreement rates for 19 doubly-reviewed items ranged from 0 to 47%, with an overall average of 10.1% (details, Supplement Table 5). Initial disagreement was greater for assessments of the credibility and clinical importance of claimed HTE. Possibly because assessment of credibility and clinical importance were added near the end of data collection and without additional training, initial disagreement was more common for these items, although generally this was easily resolved upon discussion.

Risk Models. Concordance with eight Statement criteria related to risk modeling was above 60% for 5 of 8 (Supplement Table 3). Only 52% (25) of the 48 reports with positive overall findings included a risk model. Fourteen of 31 risk modeling analyses used an external prediction model. Half of reports presented risk model scores by treatment arm. All but one presented absolute treatment effects by level of risk and most reported relative treatment effects as well.

Study authors claimed findings of HTE in 23 of 31 risk modeling analyses (Figure 2, Supplement Table 8). For 13 reports, HTE was found on the absolute but not the relative scale (i.e., risk magnification). For the remaining 10, relative treatment effects also appeared to vary across levels of baseline risk. In five, (49,61,79,91,96) relative treatment effects were greater in individuals at higher risk for experiencing trial outcomes. Relative treatment benefit was confined to individuals in the middle of the risk distribution in three reports (47,65,69) or to those at lowest risk in two (59,84).

Effect Models. Concordance was low for 3 of the 6 Statement criteria among effect modeling reports (Supplement Table 4). Only 6 of 41 (52,60,84,86,108,109) restricted analyses to small numbers of covariates with strong prior evidence for effect modification. Most explored many candidate effect modifiers with little prior evidence. Only nine (51,52,53,55,70,80,92,93,108) applied effect model findings to external datasets for validation. Authors claimed HTE in 31 of the 41 effect model reports (Figure 2, Supplement Table 8). Thirty presented evidence for absolute treatment effect differences across subgroups; nine presented evidence for relative effect differences. In 14 of the 31, authors heeded recommendations to report model performance metrics that evaluate prediction of individual treatment effects rather than prediction of risk for the outcome.

Assessment for Credibility of HTE. Most reports, whether of risk or effect modeling, claimed to have identified HTE (Figure 2), but as expected, risk models were more likely to be scored as credible when ICEMAN criteria were applied. Detailed scores for the 51 reports claiming HTE are given in Supplement Table 8. Twelve of 13 risk modeling analyses claiming risk magnification and eight of ten that found HTE on a

relative scale were scored as credible. By contrast, findings of HTE in 31 effect modeling reports were judged credible in only 10. Most effect modeling reports explored many variables with little prior evidence for effect modification. Those employing data-driven machine-learning algorithms also allowed the data to guide cut-point selection for continuous covariates. Among the 10 effect modeling reports judged to present credible HTE, nine (51,52,55,70,80,92,93,108,111) validated model predictions of individual treatment effects in independent cohorts, usually another RCT. Three of the ten (52,99,108) also met ICEMAN criteria (and PATH recommendations) by restricting analyses to a very small number of candidate effect modifiers with strong prior evidence.

Assessment for Clinically Important HTE. Reviewers judged findings from 24 of 30 reports with credible HTE to be clinically important (Figure 2). Details for these are 24, including rationales for classification as clinically important, are summarized in Table 2. In 19, RCT findings had suggested a moderate (n=6) or strong (n=13) benefit of one treatment vs. another. Yet, predictive modeling identified subgroups representing 5 to 67% of the trial population for whom no benefit or possible net harm would be expected from that treatment. In five, overall results suggested no benefit, but predictive modeling identified subgroups representing 25-67% of participants who did appear to benefit from one treatment vs. another. Findings of six reports with credible HTE were judged not clinically important because the heterogeneity, though credible, did not span a threshold suggesting differing treatment choices; (95,112) or because of conflicting findings across outcomes, (75,88) failure to add clinical value to previous risk-based selection

strategies, (79) or concurrence with authors on the need for additional investigation, possibly testing additional effect modifiers. (55)

Discussion

Evidence-based medicine has historically encouraged clinicians and patients to rely on average treatment effects from RCTs to support individual decision-making, (23,115) despite recognizing limitations with this approach. In the four and a half years following publication of the PATH Statement, a steadily growing number of publications across a wide range of clinical areas has employed predictive modeling to examine possible HTE in RCT results. Among the 65 reports we identified, more than a third found HTE that was both credible and clinically important, suggesting that patients and clinicians could often do better than relying solely on average effects.

Consistent with the PATH Statement rationale and ICEMAN criteria, risk modeling was more likely than effect modeling to produce findings of credible HTE because of its relative simplicity. Importantly, HTE was not always confined to the absolute scale (risk magnification). In nine reports, credible and important HTE was identified for relative as well as absolute treatment effects across baseline risk. In five, (49,54,61,79,96) relative as well as absolute effects were greater for persons at higher predicted risk. In two, (59,84) persons at low risk benefited while high-risk patients may have been harmed by the same treatment; and in two, (47,69) maximal benefit was found for those in the mid-range of risk. This U-shaped, or “sweet spot,” pattern (47) is clinically intuitive and has also been observed elsewhere. (116)

These findings demonstrate that simple assumptions of risk magnification are not well-founded and illustrate the potential value of routine risk modeling of RCT results. They may offer clinical insights about specific effect modifiers. Traits incorporated into risk scores because they are strong predictors of study outcomes may sometimes also be treatment effect modifiers, either directly or as proxies for unmeasured attributes. In an RCT comparing therapeutic-dose heparin with usual thromboprophylaxis for patients hospitalized with COVID-19, (84) respiratory status at baseline was the most potent predictor of clinical outcomes but was also found to be a strong modifier of heparin treatment effect in initial subgroup analyses. Only patients with better baseline respiratory status benefited from heparin. When this trait was incorporated into a risk model, only patients with lower risk scores appeared to benefit. In three RCTs (47,59,69) where incidence of study outcomes was particularly high (range 27-61%), no benefit was observed in the highest stratum of predicted risk. For such extremely high-risk individuals, risk prediction models likely included attributes reflecting irreversible disease or competing causes of the outcome that would make treatment futile.

The increasing use of effect modeling, particularly exploratory machine-learning approaches, suggests enthusiasm for moving beyond risk stratification to more flexible estimation of individualized treatment effects. Several authors (58,76,88) expressed concerns that although risk scores can create patient subgroups well-matched on risk, subgroup members may still be heterogenous for the specific characteristics that contributed to their risk scores and therefore potentially heterogeneous in their responses to treatment. Comparative advantages and disadvantages of risk versus effect modeling remain incompletely understood. Seven reports (57,81,83,91,105,

109,111) presented both risk and effect modeling of the same RCT data. Effect modeling added new insights to risk modeling in only one instance. (111) In this IPDMA of eight RCTs of corticosteroids for community-acquired pneumonia, effect modeling identified a single powerful relative treatment effect modifier, c-reactive protein, a variable that was not included in the external risk model. Although both models found credible HTE, the effect model performed better in external validation. More generally, the existence of strong effect modifiers, whether known in advance or not, is a likely pre-requisite for finding that an effect model improves on risk modeling.

Nearly all effect modeling reports followed Statement suggestions to use shrinkage methods and internal validation strategies to reduce over-fitting (Supplement Table 4). Nevertheless, inconsistencies in several reports illustrate the persistent challenges of false positive signals of HTE in exploratory effect modeling and underscore the need for external validation. For example, two reports (55,82) applied causal forest algorithms to data from the SPRINT and Action to Control Cardiovascular Risk in Diabetes (ACCORD) trials evaluating intensive systolic blood pressure control. One (82) found evidence of HTE, the other did not. In two reports (60,78) from a trial of dabigatran vs. warfarin for stroke prevention in atrial fibrillation, one (60) suggested interactions of three covariates with treatment choice and significant HTE; the second, using four machine-learning algorithms applied to the same RCT data, found no evidence for HTE. Two reports (62,110) compared results of multiple machine-learning algorithms, finding inconsistent evidence for HTE between algorithms and even within algorithms when random initiation seeds were altered.(62) Two reports (92,111)

compared regression-based methods with machine-learning algorithms in effect modeling, both finding that regression models performed better in external validations.

The inconsistency of machine-learning approaches to effect modeling has also been addressed by others (117). Most effect models reviewed here explored large numbers of candidate treatment interactions despite relatively modest numbers of outcome events. Specification of best practices in this emerging area is beyond the scope of the present review, and much remains to be learned concerning sample size needs and optimal approaches to internal validation.

Meanwhile, the value of external validation illustrated here highlights the importance of making data from completed RCTs available and of creating large, well-characterized real-world cohorts with treatment and covariate data for validation and extending HTE findings to populations with differing patterns of risk and treatment. (118) The same cohorts could support development of new, more representative risk prediction models.

Given the frequent utility of risk modeling, sponsors of new RCTs should consider in advance whether appropriate external risk models exist and plan for collection of baseline data needed for estimating individual risk. Within the past decade, editorial guidelines for reporting positive RCTs findings have come to require presentation of absolute as well as relative measures of overall treatment effect because of their greater relevance in clinical decision-making. (119-121) We suggest that reports of positive RCTs could be further enhanced by requiring that treatment effects, in both relative and absolute terms, be presented in relation to baseline risk.

Ultimately, it will remain critical to demonstrate the safety and effectiveness of any predictive model when employed for personalizing treatment choices in real world populations.

Limitations

The search strategy would not have captured predictive modeling reports that did not cite the PATH Statement. We conducted a broad search for such reports using title/abstract words “randomized” plus “heterogeneity of treatment effects” for the same time period. This strategy yielded 60 publications, but only seven were predictive models in RCTs, 3 of which were included in our review. That this search found so few of the 65 reports we reviewed indicates that it is not a parallel approach for finding other predictive models. Nevertheless, we believe that the reports citing the PATH Statement offer a highly relevant population for assessing its influence. Even with the ICEMAN and PATH criteria for assessing credibility and clinical importance of HTE, some subjectivity remains. 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. Although effect modeling holds promise for predicting individualized treatment effects, 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.

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, et al. 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, et al. 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, et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010;340:c117.

11. 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.

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

13. 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.

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

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

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

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

21. 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.

22. 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(1):74.

23. 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.

24. Barratt A, Wyer PC, Hatala R, et al; for The Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. CMAJ. 2004;171:353-58.

25. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet 2017;390:415-23.

26. Grundy SM, Stone NJ, Bailey AL, et al. 2018 AHA/ACC/AACVPR/AAPA/ ABC/ACPM/ADA/ AGS/ APhA/ ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2019 Jun 18;139(25):e1082-e1143.

27. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ ABC/ACPM/AGS/ APhA/ASH/ ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A Report of the American College of Cardiology/American Heart Association task force on clinical practice guidelines. J Am Coll Cardiol 2018;71(19):e127–e248.

28. 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.

29. 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.

30. 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.

31. Albuquerque AM, Tramujas L, Sewanan LR, et al. 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.

32. 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.

33. 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.

34. Dianti J, McNamee JJ, Slutsky AS, et al. Determinants of Effect of Extracorporeal CO_2 Removal in Hypoxemic Respiratory Failure. NEJM Evid. 2023 May;2(5):EVIDoa2200295.

35. 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. Res Sq [Preprint]. 2023 Feb 23:rs.3.rs-2441086.

36. 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 Med. 2023 Jun 6;20(6):e1004176.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. Foy AJ, Schaefer WE, Ruzieh M, et al. Re-Analyses of 8 Historical Trials in Cardiovascular Medicine Assessing Multimorbidity Burden and Its Association with Treatment Response. Am J Med 2024;137:608-16.

44. Bertismas D, Koulouras AG, Margonis GA. The R.O.A.D. to precision Medicine. arXiv:2311.01681. Accessed at https://doi.org/10.48550/arXiv.2311.01681, accessed on Nov 14, 2024.

45. Cheng J, Levy DE, McCurley JL, et al. Differential effect by chronic disease risk: A secondary analysis of the ChooseWell 365 randomized controlled trial. Prev Med Rep. 2024;42:102736

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

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

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

49. Chalkou K, Hamza T, Benkert P, et al. Combining randomized and non-randomized data to predict heterogeneous effects of competing treatments. Res Synth Methods. 2024;15:641-56.

50. 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.

51. 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 Netw Open. 2022;5(1):e2147375.

52. 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.

53. 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. J Clin Epidemiol. 2020;125:47e56.

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

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

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

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

58. 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.

59. 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.

60. 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.

61. 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.

62. 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.

63. 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.

64. 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.

65. 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.

66. 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.

67. Di Stefano L, Ogburn EL, Ram M, et al; Pandemic Response COVID-19 Research Collaboration Platform for HCQ/CQ Pooled Analyses. Hydroxychloroquine/ Chloroquine for the Treatment of Hospitalized Patients with COVID-19: An Individual Participant Data Meta-Analysis. PLoS One. 2022;17(9):e0273526.

68. 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.

69. 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. Crit Care Med. 2022;50(3):469-479.

70. 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.

71. 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.

72. 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.

73. Chen X, Harhay MO, Tong G, et al. A Bayesian Machine-Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to a Critical Care Trial. Ann Appl Stat. 2024 Mar;18(1):350-374.

74. 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.

75. 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.

76. 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. 2024 Jun 3;31(8):945-953.

77. 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.

78. 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.

79. 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. 2024 Apr;85(4):373-381.

80. 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.

81. Xu Y, Bechler K, Callahan A, et al. Principled estimation and evaluation of treatment effect heterogeneity: A case study application to dabigatran for patients with atrial fibrillation. J Biomed Inform. 2023;143:104420.

82. 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.

83. 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.

84. 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.

85. 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.

86. 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.

87. 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;6:e2315315.

88. 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. 2024 Feb 1;35(2):216-228.

89. 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.

90. 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.

91. Harrer M, Baumeister H, Cuijpers P, et al. Predicting effects of a digital stress intervention for patients with depressive symptoms: Development and validation of meta-analytic prognostic models using individual participant data. J Consult Clin Psychol. 2024;92:226-235.

92. Venkatasubramaniamm A, Mateen BA, Shields BM, et al. Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: an application for type 2 diabetes precision medicine. BMC Med Inform Decis Mak. 2023;23:110.

93. Buell KG, Spicer AB, Casey JD, et al. Individualized Treatment Effects of Oxygen Targets in Mechanically Ventilated Critically Ill Adults. JAMA. 2024;331:1195-1204.

94. Xu E, Vanghelof J, Wang Y. Outcome risk model development for heterogeneity of treatment effect analyses: a comparison of non-parametric machine learning methods and semi-parametric statistical methods. BMC Med Res Methodol. 2024;24:158 doi: 10.1186/s12874-024-02265-8.

95. Li Y, Devonshire A, Huang B, Andorf S. Risk subgroups and intervention effects among infants at high risk for peanut allergy: A model for clinical decision making. Clin Exp Allergy. 2024;54:185-194.

96. Paules CI, Wang J, Tomashek KM, et al. A Risk Profile Using Simple Hematologic Parameters to Assess Benefits From Baricitinib in Patients Hospitalized With COVID-19: A Post Hoc Analysis of the Adaptive COVID-19 Treatment Trial-2. Ann Intern Med. 2024;177:343-52.

97. Vickers A, Vertosick E, Langsetmo L, et al. Estimating the Effect of Radical Prostatectomy: Combining Data Fr om the SPCG4 and PIVOT Randomized Trials With Contemporary Cohorts. J Urol. 2024;212:310-19.

98. Zhou Z, Jian B, Chen X, et al. Heterogeneous treatment effects of coronary artery bypass grafting in ischemic cardiomyopathy: A machine learning causal forest analysis. J Thorac Cardiovasc Surg. 2024;168:1462-71.

99. Arnold SV, Jones PG, Maron DJ, et al. Variation in Health Status With Invasive vs Conservative Management of Chronic Coronary Disease. J Am Coll Cardiol. 2024;83:1353-1366.

100. Bond MJG, van Smeden M, Degeling K et al. Predicting Benefit From FOLFOXIRI Plus Bevacizumab in Patients With Metastatic Colorectal Cancer. JCO Clin Cancer Inform. 2024; Jul:8:e2400037. doi: 10.1200/CCI.24.00037.

101. Desai RJ, Glynn RJ, Solomon SD, et al. Individualized Treatment Effect Prediction with Machine Learning — Salient Considerations. NEJM Evid. 2024;3:DOI: 10.1056/EVIDoa2300041.

102. Samuels THA, Molloy SF, Lawrence DS, et al. Personalised risk prediction tools for cryptococcal meningitis mortality to guide treatment stratification; a pooled analysis of two randomised-controlled trials. MedRxiv 2024; accessed at https://doi.org/10.1101/2024.07.10.24310212 on Nov 15, 2024.

103. Hamaya R. Application of novel technologies to cardiovascular prevention. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences. 2024; Chapter 3;53-95; accessed at https://dash.harvard.edu/handle/1/37377884 on Nov 15, 2024.

104. Afshar M, Graham Linck EJ, Spicer AB, et al. Machine Learning-Driven Analysis of Individualized Treatment Effects Comparing Buprenorphine and Naltrexone in Opioid Use Disorder Relapse Prevention. J Addict Med. 2024;18:511-519.

105. Hoogland J, Takada T, van Smeden M, et al. Prognosis and prediction of antibiotic benefit in adults with clinically diagnosed acute rhinosinusitis: an individual participant data meta-analysis. Diagn Progn Res. 2023;7:16. doi: 10.1186/s41512-023-00154-0.

106. Huang Q, Zou X, Chen Y, et al. Personalized glucose-lowering effect of chiglitazar in type 2 diabetes. iScience, 2023;26:108195. doi: 10.1016/j.isci.2023.108195.

107. Luo Y, Chalkou K, Funada S, et al. Estimating Patient-Specific Relative Benefit of Adding Biologics to Conventional Rheumatoid Arthritis Treatment: An Individual Participant Data Meta-Analysis. JAMA Netw Open. 2023; 6:e2321398. doi: 10.1001/jamanetworkopen.2023.21398.

108. Ninomiya K, Kageyama S, Shiomi H. Can Machine Learning Aid the Selection of Percutaneous vs Surgical Revascularization? J Am Coll Cardiol. 2023;82:2113-2124.

109. de Winkel J, Roozenbeek B, Dijkland SA, et al. Personalized decision-making for aneurysm treatment of aneurysmal subarachnoid hemorrhage: development and validation of a clinical prediction tool. BMC Neurol. 2024;24:65. doi: 10.1186/s12883-024-03546-x

110. Bouvier F, Peyrot E, Balendran A, et al. Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data. Stat Med. 2024;43:2043-2061.

111. Smit JM, Van Der Zee PA, Stoof SCM, et al. Predicting individualized treatment effects of corticosteroids in community-acquired-pneumonia: a data-driven analysis of randomized controlled trials. Lancet Respiratory Medicine 2025; Jan 15, in press.

112. Burger PM. Residual risk in vascular disease and heart failure: Risk factors and individualized prevention. Doctoral dissertation, University of Utrect, NL. 2024; accessed at https://dspace.library.uu.nl/handle/1874/437330 on Nov 15, 2024.

113. 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.

114. Dahabreh IJ, Kazi DS. Toward Personalizing Care: Assessing Heterogeneity of Treatment Effects in Randomized Trials. JAMA. 2023;329:1063-65.

115. 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.

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

117. Lipkovich I, Svensson D, Ratitch B, et al. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. Clin Trials. 2023;20(4):380-393.

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

119. 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.

120. 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.

121. 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. Characteristics of the 70 analyses (65* reports) of predictive models for possible heterogeneity of treatment effect

	Risk Modeling Analyses (n=31)	Effect Modeling Analyses (n=41)
Publication Year		
2020	2	3
2021	4	8
2022	7	7
2023	9	15
2024 (through July 5)	9	8
Initial Publication Status		
Peer-reviewed	24	34
Pre-print archive[†]	3	7
Dissertation	4	1
Data Source		
Re-analyses of Single RCT	16	26
Initial Analyses of RCT(s)	2	0
IPDMA of 2 or more RCTs	13	15
Comparative Effectiveness Research? [‡]		
Yes	13	25
No	18	16
Overall RCT Results [§]		
Null (no overall difference)	6	12
Modest effect	7	13
Strong effect	18	16
Sample Size - HTE Analyses		
Range (min - max)	574 – 330,460	200 – 26,877
Median	1,907	2,294
Interquartile Range (25th - 75th)	999 – 3,740	1,250 – 8,828
For Risk Models Only (n=30)		
Type of Risk Model Used		
External model applied	14	
Internal model developed	17	**16**
For Effect Models Only (n=41)		
Number of Potential Interactions Tested		
Range (min - max)	-	2 – 58
Median	-	17
Interquartile Range (25th - 75th)	-	10 – 22

Type(s) of Effect Models Employed		-
Regression Methods (n=13)[‖]		
"Conventional" Regression[¶]	-	6
Regression with penalization to avoid over-fitting**	-	7
Machine-Learning Algorithms (n=28)[††]		
Causal Forests	-	13
Other tree-based Methods[‡‡]	-	10
Meta-Learner Methods	-	9

*Seven reports (refs 60,84,86,91,105,109,111) presented both risk and effect modeling analyses.

[†] Seven of 10 reports (refs 49,70,73,76,81,88,111) originally identified through pre-print archives have subsequently been published in peer-reviewed journals. Three (refs 58,89,102) have not appeared as peer-reviewed publications.

[‡] Comparative effectiveness was defined as comparison of two or more alternative, active interventions. (for more detail, see coding instructions Supplement Table 4).

[§] Null: overall treatment effect does not differ significantly from zero; Modest effect: estimated relative risk reduction is $\leq$ 20%; Strong effect: estimated relative risk reduction >20%. For continuous outcomes, a standardized mean difference significantly >0, but $\leq$ 0.8 was considered moderate; and a standardized mean difference >0.8 was considered strong. In one report (ref 78), an overall trial effect could not be determined.

[‖] These include only reports based solely on regression models without additional analyses employing more flexible machine-learning algorithms.

[¶] "Conventional regression" includes linear, logistic and Cox proportional hazards regression models that did not employ either penalization / regularization methods (e.g., LASSO, penalized ridge regression, elastic net) or cross-validation methods explicitly intended to reduce over-fitting.

** These include regression models that incorporated penalization /regularization methods or cross-validation methods explicitly intended to reduce over-fitting (or both)

[††] Because some reports featured more than one machine-learning algorithm, the rows below are not mutually exclusive

[‡‡] Includes examples of model-based recursive partitioning, gradient-boosted regression trees, random forest, and Bayesian additive regression trees (BART)

Table 2. Studies found to have credible and clinically important heterogeneity of treatment effects (HTE)

Reference	Clinical Condition	Outcomes	Randomized Intervention (s)	Overall RCT Findings	Methods	Type of HTE Identified	Clinically Important Subgroup Differences
RISK MODELS (n=15)							
Redelmeier et al.[46,47]	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 heterogeneity, p<0.001 for multiplicative interaction	Despite a strong overall benefit of ICD implantation, the mortality benefit was largely confined to patients in midrange of risk (3rd and 4th quintiles), i.e., a "sweet spot." Intervention associated with high costs, patient discomfort.
Chalkou et al.[48,49]	Multiple sclerosis (MS)	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 RCTs, using network meta-analysis; external risk model	Probable relative as well as absolute effect heterogeneity; test for multiplicative interaction not reported	Despite findings of strong overall benefit of natalizumab, 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).
Kumar et al.[54]	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 heterogeneity; test for multiplicative interaction not reported	Despite a strong overall benefit of apixaban, patients with a baseline risk for VTE < 8% (67% of the trial population) derived no benefit, but experienced an excess of overall bleeding events on treatment with apixaban
Bress et al.[60]	Systolic hy-pertension and increased cardiovas-cular 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 heterogeneity (risk magnification)	Despite the strong overall benefit of intensive SBP control, patients in lowest risk quartile would expect very little benefit from intensive control, despite increased costs, burden and adverse effects.

Reference	Clinical Condition	Outcome(s)	Randomized Interventions	Overall RCT Findings	Methods	Type of HTE Identified	Clinically Important Subgroup Differences
Rysavy et al.[59]	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 heterogeneity, p=0.03 for multiplicative interaction	In the face of a weak overall benefit of Vitamin A therapy, predictive modeling found a strong benefit in the 50% of infants at lower predicted risk; no suggestion of benefit among infants in highest quarter of predicted risk.
Kent et al.[61]	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 RCTs; external risk model	Relative as well as absolute effect heterogeneity, p=0.003 for multiplicative interaction	Despite a strong overall benefit, analyses Identified a subgroup comprising 15% of trial population unlikely to have had a PFO-related stroke. This subgroup received no benefit from PFO-closure and were at higher risk of procedure-related complications.
Taylor et al.[69]	Hospitalized Patients with Sepsis	30-day mortality and readmission	Nurse-navigator led Sepsis Transition and Recovery (STAR) intervention vs usual care	Weak benefit in favor of the intervention (aOR=0.80)	Single RCT re-analysis; external risk model	Relative as well as absolute effect heterogeneity; 95% confidence intervals for quartile-specific OR's do not overlap	Despite findings of a modest overall benefit of the program, the benefit was confined to patients in the midrange of predicted risk (middle two quartiles) i.e., "sweet spot"). Intervention is associated with greater costs.
Gencer et al.[71]	Atrial Fibrillation	Net composite of stroke/ systemic embolism, major bleed, all-cause death	Lower (LDER) vs higher (HDER) dose regimens of edoxaban vs warfarin	Weak benefit for either dose of edoxaban vs warfarin (HR=0.83, 0.89 for LDER,HDER respectively)	Single RCT re-analysis; Risk stratification based on count of number of high-risk features	Absolute effect heterogeneity (p=0.001) but probably not relative effect heterogeneity (p=0.065) (risk magnification)	With a weak overall effect, the absolute benefits of either dose of edoxaban vs warfarin were shown to be negligible in those with 0-1 risk factors, but to increase for both stroke/embolism and major bleeding endpoints as number of risk factors increases. May not justify switching in low-risk patients well managed on warfarin.

Reference	Clinical Condition	Outcome(s)	Randomized Interventions	Overall RCT Findings	Methods	Type of HTE Identified	Clinically Important Subgroup Differences
Trinks-Roerdink et al.[82]	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 heterogeneity (p=0.93 for multiplicative interaction) (risk magnification)	Despite a strong overall benefit of the program, benefits were shown to be predominantly in persons in highest 25% of risk distribution; minimal benefit with increased increased costs in lower 75%.
Goligher et al.[84]	Hospitalized COVID-19 infection	Organ-support free days; hospital survival	Therapeutic-dose heparin vs usual pharmacologic thromboprophy-laixis	Null (No overall benefit in population (OR for benefit 1.05)	IPDMA – 3 trials; re-analysis; subgroup analyses, internal risk model and effect model)	Apparent qualitative interaction on both relative and absolute scales; statistical significance not tested for risk model	Although no overall benefit was observed, a clear benefit was shown for the 60% of persons in lower 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); note: similar HTE was suggested in subgroup analyses and in an effect model
Xu et al.[94]	Healthy people 70+ yrs old;	Composite of death, dementia or disability	Low-dose (100mg) aspirin vs placebo	Null (HR=1.01, 95%CI: 0.92-1.11)	Single RCT re-analysis, external risk model	Absolute effect heterogeneity (p=0.03), but no relative effect heterogeneity (risk magnification)	Despite a null overall effect, risk model showed that highest risk 20% of patients benefitted (ARR = 15.1%; 95% CI 4.0–26.3%), with little suggestion of any benefit in remaining 80%.
Paules et al.[96]	Patients hospitalized with Covid-19	28-day mortality; progression to mechanical ventilation or death; recovery rate.	Patients hospitalized with Covid 19	Baricitinab + remdesivir vs. remdesivir alone	Strong Benefit of combination for each outcome. (HR for death, 0.65; 95% CI, 0.39 to 1.09)	Single RCT re-analysis, external risk model	Clear heterogeneity on absolute scale for all 3 outcomes (risk magnification); also, possible relative effect heterogeneity for the time-to-recovery outcome, with HR for for the highest risk quartile significantly greater than for lowest risk quartile	A strong overall benefit of baricitinib in the trial was shown to be due to a very strong benefit on absolute and relative scales in the highest risk quartile, and likely the 2nd highest risk quartile as well; with little apparent effect in the lower 50% of predicted risk.

Reference	Clinical Condition	Outcome(s)	Randomized Interventions	Overall RCT Findings	Methods	Type of HTE Identified	Clinically Important Subgroup Differences
Vickers et al.[97]	Localized prostate cancer	15-year prostate cancer mortality	Radical prostatectomy vs conservative care	Strong benefit of surgery (HR=0.55)	Separate RCT re-analyses from two trials, external risk models	Established that HR's were constant across levels of baseline risk (p-values for interaction on relative scale >0.20 in both trials); therefore: risk magnification	Despite a strong overall relative benefit,the authors show that absolute benefit of radical prostatectomy in the lowest risk quartile of current cohorts would be approximately a 0.2% reduction in 15-year mortality. For higher risk subgroups, benefit is substantial. This is useful information for shared decision-making
De Winkel et al.[109]	Acute subarachnoid hemorrhage	Functional status at 2 mos; re-treatment or re-bleed by 10 yrs.	Endovascular coiling vs intracranial clip	Strong benefit of coiling for good functional status (OR for benefit: 1.5, 95% CI 1.3, 1.7); but a strong benefit of clipping for 10-yr treatment failure: (HR: 0.30, 95% CI 0.2, 0.4)	Single RCT re-analysis; ordinal regression for functional status; proportional hazards model for treatment failure; began as exploratory effect modeling but reduced to risk modeling	Established strong risk magnification favoring intracranial clip for outcome of re-bleed or re-bleed. Modest differences in absolute benefits of coiling across range of risk for poor functional outcome	Despite competing overall findings for the two outcomes, indicating need for shared decision-making, a subgroup of 6% were predicted to have a strong benefit of clipping for the re-bleed/retreatment outcome with very low risk and negligible (<2%) differences between treatments for the functional status outcome.
Smit et al.[111]	Community-acquired pneumonia	30-day mortality	Corticosteroids vs placebo	Strong benefit of corticosteroids (OR 0.72; 95% CI 0.56-0.94)	IPDMA of 6 trials, external risk models; also examined an effect model (see below)	No evidence of treatment-risk interactions for relative effects, but moderate risk magnification	Despite a strong overall benefit of corticosteroids, the risk model identified 36% of low-risk patients with almost no predicted benefit (and increased risks corticosteroid-related risks for re-hospitalization and hyperglycemia)
Reference	**Clinical**	**Outcome(s)**	**Randomized**	**Overall RCT**	**Methods**	**Type of HTE**	**Clinically Important Subgroup**

	Condition		Interventions	Findings		Identified	Differences
EFFECT MODELS (n=9)							
Park et al.[51]	Hospitalized Covid Infection without mechanical ventilation	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	Despite null findings in overall trial population, predictive modeling identified 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.
Takahashi et al.[52]	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=0.84 for 10-year mortality; HR=0.78 for 5-year adverse cardiac events)	Single RCT-reanalysis; Cox proportional hazards regression; external validation in IPDMA of 3 trials	Multiplicative interactions of treatment choice with two pre-specified effect modifiers.	Predictive 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.
Dennis et al.[70]	Type 2 diabetes	Hb A1c at 6 mos post-initiation	Initiation of SGLT-2 inhibitors vs DPP-4 inhibitor as add-on therapy	Weak benefit in favor of SGLT-2 inhibitor: mean difference in Hb A1c: 1.2 mmol/mol	Model built in large obs. cohort; linear regression; external validation in multiple RCTs.	Clear subgroup heterogeneity in absolute differences in treatment effects on Hb A1c control.	Predictive model identifies a large subgroup, comprising 60% of trial participants, with strong benefit 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.
Seitz et al.[80]	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).	Despite null findings in the overall trial population, predictive modeling 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 no treatment effect differences in the remainder.
Reference	**Clinical Condition**	**Outcome(s)**	**Randomized Interventions**	**Overall RCT Findings**	**Methods**	**Type of HTE Identified**	**Clinically Important Subgroup Differences**

Venkatasu-bramaniam et al.[92]	Diabetes, type 2	Hb A1c reduction at 6 mos.	SGLT-2 inhibitors vs DPP-4 inhibitors as add-on therapy	A very modest benefit of SGLT-2 inhibitors (1.6 mmol/mol greater Hb A1c reduction at 6 mos)	IPDMA of 2 trials; penalized linear regression and causal forest models, external validation in large observational cohort (n=18,741)	Both penalized regression and causal forest models indicated significant HTE (p=0.005 for causal forest); significance testing not reported for regression model, but benefit prediction in external cohort was excellent	Regression model reliably identified a subgroup of 87% of validation cohort who did better if initiated on an SGLT-2 inhibitor, including 30% of all patients who were predicted to have a >5 mmol/mol larger reduction in Hb A1c if started on the SLGT-2 inhibitor. For 13% of patients, the Hb A1c reduction was predicted to be better with DPP-4 inhibitors.
Buell et al.[93]	Patients requiring mechanical ventilation	28-day mortality	Lower vs higher 02 saturation targets	Null findings in multiple trials of moderate size.	Single RCT re-analysis; ensemble machine-learning effect models (RBoost); external validation in a second RCT	Both relative and absolute HTE (p=0.02 for interaction) shown across tertiles of predicted benefit in the validation trial	Null overall findings in 7 prior RCTs comparing lower vs higher 02 saturation targets trials; these validated analyses identified subgroups of patients who clearly benefited from either lower (33% of all patients) or higher O2 saturation targets and also identified the most critical individual characteristics for predicting benefit.
Arnold et al.[99]	Patients with chronic coronary artery disease	Angina symptoms at 12 months follow-up	Coronary revascularization vs medical therapy	A modest benefit of revascularizatio n for angina symptoms	Single RCT re-analysis; ordinal proportional odds regression effect model	Strong absolute HTE, (<0.01) for a single baseline covariate – symptom severity.	Weak overall benefit of revascularization; effect model identified a single strong effect modifier and subgroups of 75% in whom the predicted benefit was strong, and 25% predicted to experience little if any benefit.
Ninomiya et al.[108]	Patients with de-novo three-vessel or left main coronary artery disease 3-vessel	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=0.84 for 10-year mortality; HR=0.78 for 5-year adverse cardiac events)	Single RCT re-analysis; pro-portional hazards regression effect model with two pre-specified interactions; ex-ternal validation in observational cohort	Strong absolute and relative scale HTE (p<0.005 for two covariates)	Despite relatively strong overall benefit of CABG, effect model identifies 25% of patients who would have better predicted outcomes with PCI, 50% with very strong benefit of CABG, and 25% for whom there is little predicted difference in benefits.
Reference	**Clinical Condition**	**Outcome(s)**	**Randomized Intervention**	**Overall RCT Findings**	**Methods**	**Type of HTE Identified**	**Clinically Important Subgroup Differences**
Smit et	Community-	30-day	Corticosteroids	Strong benefit	IPDMA of 6 trials;	Strong HTE driven	Despite a relatively strong overall

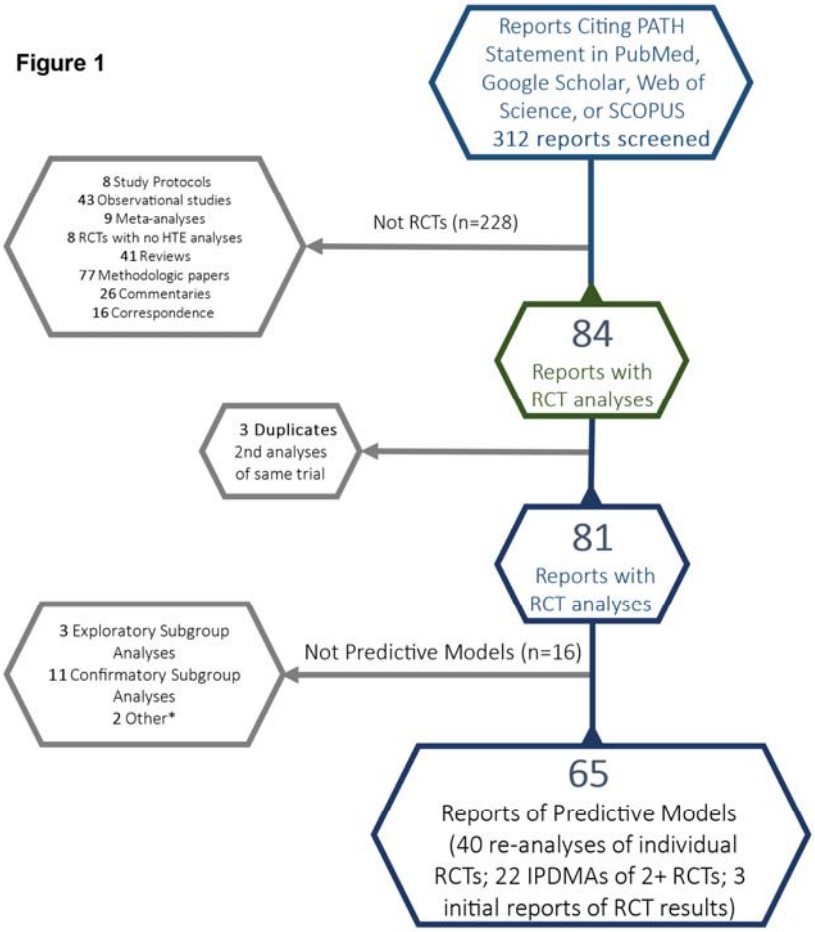
al.[111]	acquired pneumonia	mortality	vs placebo	of corticosteroids (OR 0.72; 95% CI 0.56-0.94)	penalized logistic-regression based effect models; external validation in 2 addition RCTs.	primarily by interaction with a single baseline covariate, the C-reactive protein level. (p=0.05)	benefit of corticosteroids, the effect model identified nearly half of patients, with lower levels of CRP, with no predicted benefit from corticosteroids, and experienced adverse effects including re-hospitalization and hyperglycemia

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.

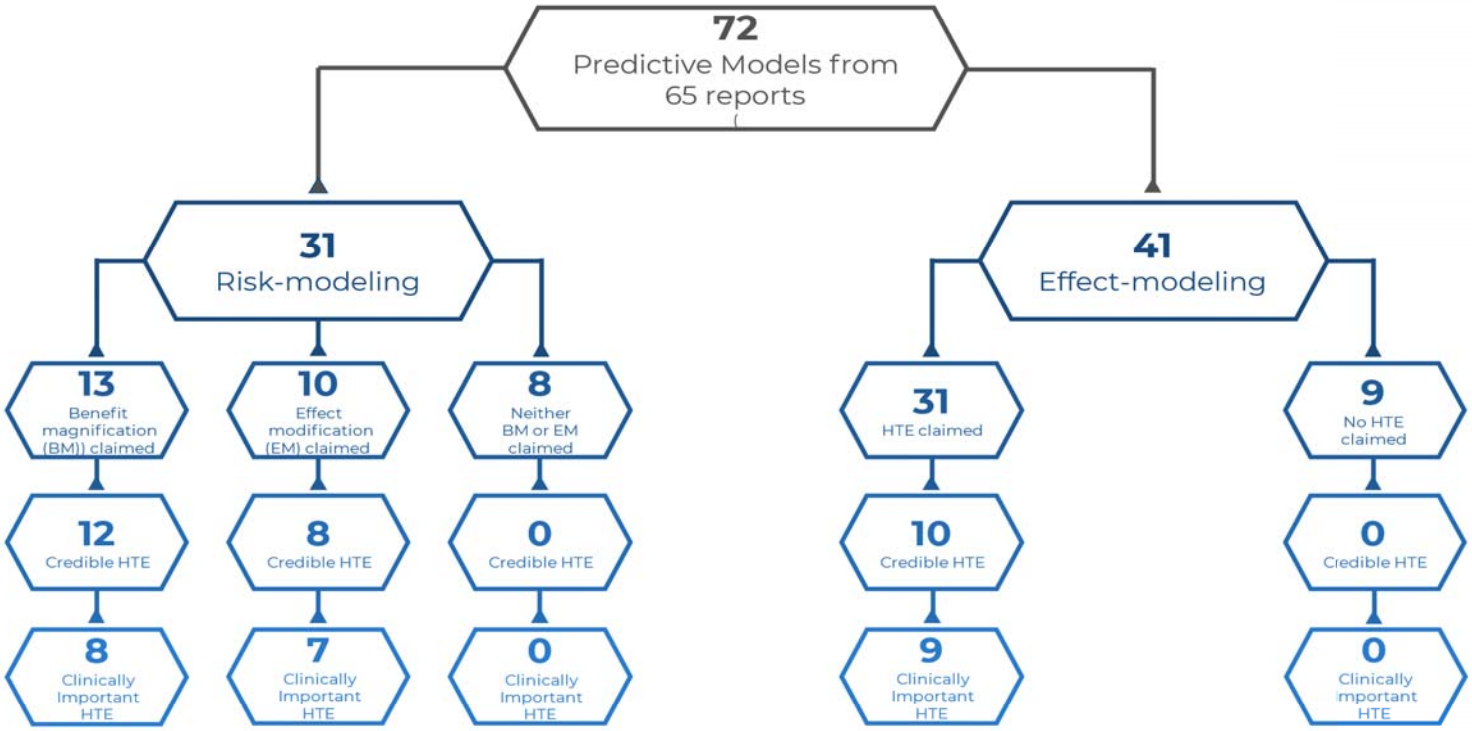
Figure 2. Adjudicated results of review of all eligible reports for type of predictive modeling (risk or effect), for claims by authors of heterogeneity of treatment effects (HTE), for credibility of HTE (using adapted ICEMAN criteria), and for clinical importance of HTE found to be credible.

Figure 1



*Details of these 16 studies are presented in Supplement Table 1b

Figure 2


72
Predictive Models from 65 reports
31
Risk-modeling
41
Effect-modeling
13
Benefit magnification (BM)) claimed
10
Effect modification (EM) claimed
8
Neither BM or EM claimed
31
HTE claimed
9
No HTE claimed
12
Credible HTE
8
Credible HTE
0
Credible HTE
10
Credible HTE
0
Credible HTE
8
Clinically Important HTE
7
Clinically Important HTE
0
Clinically Important HTE
9
Clinically Important HTE
0
Clinically important HTE