Assessing heterogeneity of treatment effect in real-world data

Segal, J.B.; Varadhan, R.; Groenwold, R.H.H.; Henderson, N.C.; Li, X.J.; Nomura, K.; ... ; Burcu, M.

Assessing Heterogeneity of Treatment Effect in Real-World Data

Jodi B. Segal, MD, MPH*; Ravi Varadhan, PhD, PhD*; Rolf H.H. Groenwold, MD, PhD; Nicholas C. Henderson, PhD; Xiaojuan Li, PhD; Kaori Nomura, MPH, PhD; Sigal Kaplan, PhD; Shirin Ardeshirrouhanifard, PharmD, PhD; James Heyward, MHS; Fredrik Nyberg, PhD; and Mehmet Burcu, PhD

Increasing availability of real-world data (RWD) generated from patient care enables the generation of evidence to inform clinical decisions for subpopulations of patients and perhaps even individuals. There is growing opportunity to identify important heterogeneity of treatment effects (HTE) in these subgroups. Thus, HTE is relevant to all with interest in patients' responses to interventions, including regulators who must make decisions about products when signals of harms arise postapproval and payers who make coverage decisions based on expected net benefit to their beneficiaries. Prior work discussed HTE in randomized studies. Here, we address methodological considerations when investigating HTE in observational studies. We propose 4 primary goals of HTE analyses and the corresponding approaches in the context of RWD: to confirm subgroup effects, to describe the magnitude of HTE, to discover clinically important subgroups, and to predict individual effects. We discuss other possible goals including exploring prognostic score– and propensity score–based treatment effects, and testing the transportability of trial results to populations different from trial participants. Finally, we outline methodological needs for enhancing real-world HTE analysis.

Ann Intern Med. 2023;176:536-544. doi:10.7326/M22-1510 **Annals.org**
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 21 March 2023.
* Drs. Segal and Varadhan contributed equally to the work.

The widespread availability of rich, real-world data (RWD) generated from patient care provides increased opportunities to generate evidence to inform clinical decisions for subpopulations of patients, and perhaps even for individuals (1, 2). Heterogeneity of treatment effects (HTE) describes how treatment effect varies across patients. Although there is considerable literature on HTE among patients enrolled in randomized clinical trials (RCTs) (3–10), the assessment of HTE in RWD is a newer challenge.

We first consider RCTs. In trials, the treatment effect is most commonly reported as an overall treatment effect, a comparison of the average response between 2 treatment groups; RCTs often report estimates for patient subgroups that are defined by single characteristics, sequentially, such as men versus women and older versus younger patients. A recent advance in reporting HTE in RCTs came with the Predictive Approaches to Treatment effect Heterogeneity (PATH) statement, published in 2020 (3), which provided recommendations on 2 approaches with trial data. One uses a multivariable model to predict the risk for the outcome of interest for the trial participants, with treatment effects reported within strata of prognostic risk. The second approach estimates the treatment effect among trial participants with models that include interactions between treatment and baseline covariates; this supports estimation of treatment effects that vary by patient characteristics.

Another important development is guidance from the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) report (7), which assists in interpretation of treatment effect differences in RCTs across subpopulations, that is, when there is effect modification. The core questions are whether prior evidence supports effect modification, was the direction of the effect correctly hypothesized a priori, did the test for interaction exclude a chance finding, and did investigators test only a few effect modifiers and avoid arbitrary cut points of the effect modifier? Both PATH and ICEMAN are important advances for appropriate interpretation of HTE in RCTs; it is unclear whether these recommendations are sufficient for observational studies using RWD, where treatments are not randomly assigned and there may be greater variation in treatment response than in RCTs.

When designing an observational study, investigators make similar decisions as when designing a trial: they specify eligibility to define a population at a time zero, choose an exposure and comparator of interest, assess the outcomes after time zero, and analyze all persons in the population (11, 12). Thus, PATH and ICEMAN recommendations should be relevant. However, the greater heterogeneity among real-world patients compared with trial participants creates opportunities to generate meaningful evidence for more personalized practice decisions. Also, RWD introduces the challenge of the treatment effect being confounded with the treatment uptake mechanism (that is, confounding by indication). Thus, the analysis of RWD presents both opportunities and challenges not addressed in PATH and ICEMAN.

What Are the Primary Sources of HTE?

By definition, HTE is the nonrandom variability in the direction or magnitude of individual treatment effects (13) (**Appendix Table**, available at Annals.org). This variability may be due to intrinsic biological characteristics of treated persons (genetics, clinical conditions), extrinsic environmental factors (diet, pollution), and behaviors (adherence to treatment). In addition, heterogeneity arises when there are differences in treatment access or delivery, concomitant therapies, clinician expertise, or site features. Furthermore, patients are treated by clinicians who are nested within hospitals nested within health systems; each level of nesting can contribute to variation in the observed responses to treatment.

What Motivates HTE Assessment?

For nearly every treatment, effects should be expected to differ across individuals. Rarely, a drug or vaccine is

known from preapproval trials to be exceedingly beneficial for all subgroups of recipients (14). In most situations, however, one cannot assume the absence of HTE. One reason that many regulatory authorities require postmarketing research is the likelihood of treatment risks in subpopulations that were not detected during premarket studies (15). For a clinical decision, information about the *net* benefit is most valuable. Net benefit that incorporates both benefits and harms attributable to the intervention is most valuable to clinical decision making (16, 17). Because both benefits and harms can be heterogeneous across treated persons, net clinical benefit can be heterogenous as well (18).

What Are the Key Considerations When Using RWD That Must Proceed Before Any Evaluation of HTE?

Evaluation of HTE necessarily comes *after* one has generated a valid estimate of the overall effect across all patients. There is well-described guidance for conducting reproducible real-world evidence studies of the effectiveness and safety of medical therapies (19–21). Investigators using RWD will typically make design choices that allow assessment of the effectiveness of an intervention, which may differ from its efficacy in a tightly controlled setting. Investigators will choose an appropriate approach to addressing missing data, and will address issues of selection bias when reporting results. Investigators will also consider the risk of measurement biases, assess the sufficiency of information about potential confounders, and address potential confounding using stratification, adjustment, weighting, or matching.

Yet, even if the main effect does not differ significantly or clinically between treatment groups, investigation of HTE should still proceed. The null effect could be due to 1 subgroup of persons responding strongly to 1 therapy and another responding strongly to the comparator. Without exploring HTE, valuable information for decision making is missed.

Which Effect Scale Should Be Used When Evaluating and Reporting HTE?

In comparative effectiveness research, there are 2 primary ways to compare outcomes across treatment groups: a relative comparison (ratio) or an absolute comparison (difference). Whether HTE should be assessed on a *relative* or on an *absolute* scale, depends on the purpose. An advantage of *absolute* measures is that the effect of the treatment on the subgroup can be described directly; interpretation of *relative* measures requires knowing the baseline risk for the outcome of interest without treatment (22) (Figure 1).

Some have proposed that both multiplicative (relative) and additive (absolute) interactions should be reported (23, 24). Methodologists supported by the Patient-Centered Outcomes Research Institute (PCORI) advised that: 1) HTE on the additive (absolute) scale is most interpretable to guide clinical decisions, as heterogeneity reported on the multiplicative (relative) scale may obscure the magnitude or direction of an important interaction; 2) the additive (absolute) scale may give clues about interactions that are likely to be etiologic, although etiologic considerations may not be relevant to patients' treatment choices; and 3) statistical modeling need not, necessarily, be conducted on the same scale as that with which results are communicated (25). The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist for reporting observational studies urges that estimates of relative risk be translated to absolute risk estimates across a meaningful time period (26).

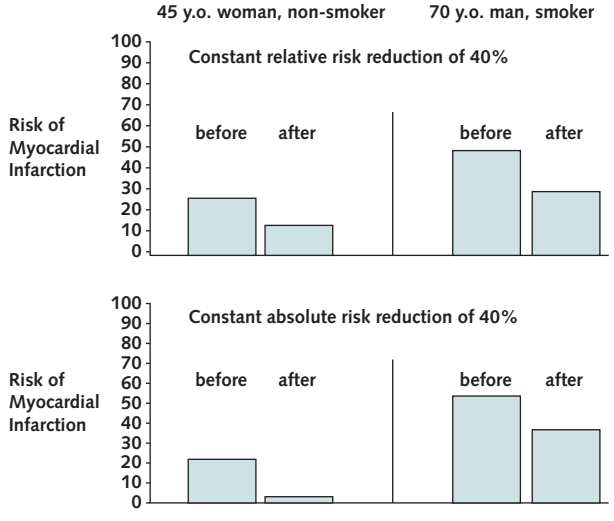
What Are the Different Objectives of Conducting HTE Analyses?

Any study may have several primary and secondary objectives regarding the assessment of HTE. We strongly encourage investigators to specify the goal in a protocol or study plan. The text that follows presents these objectives (in no essential order).

Objective 1: Confirm Subgroup HTE

When there is a signal of possible HTE in a clinical trial—perhaps in the confirmatory trials regulators might require for drug approval, or when passive surveillance systems (like the U.S. Food and Drug Administration [FDA] Adverse Event Reporting System) suggest possible harm in a subgroup—investigators should answer the hypothesis-driven question of whether a subgroup responds differently. A key requirement is specification of the effect estimate that will answer the study's clinical question—the

Figure 1. Risk for myocardial infarction.



y.o. = year old. **Top.** Assume a lipid-lowering drug reduces the risk for the myocardial infarction by a *relative* 40%, regardless of a patient's baseline risk, that is, no heterogeneity of treatment effects on the relative scale. Then, in older adults with a baseline risk of 50%, the absolute risk reduction will be 20%. In middle-aged adults with a baseline risk of 25%, the absolute risk reduction will be only 10%. **Bottom.** If the effect of a different lipid-lowering drug is homogeneous on an *absolute* scale, meaning that the treatment will lead to the same absolute risk reduction in all patients (15%), the relative effect will be larger for patients with a lower baseline risk for the outcome.

estimand. In a study focused on confirming HTE across subgroups, the estimand is generally a subgroup-specific treatment effect. The goal is to test whether the treatment effect in any subgroup is different from the overall treatment effect. An example is the studies that sought to confirm the risk for fungal infections attributable to use of sodium–glucose cotransporter-2 inhibitors (SGLT2 inhibitors) in subpopulations of treated patients. Pooled analyses of phase 3 trials of canagliflozin (an SGLT2 inhibitor) indicated an elevated risk for genital infection in the whole treated population, however, with earlier onset in women (27). This prompted confirmatory studies using RWD from patients exposed to SGLT2 inhibitors to better understand the effects in subgroups defined by sex in usual care settings. A cohort study using data from Ontario demonstrated a risk for genital infections associated with SGLT2 inhibitors, relative to dipeptidyl peptidase-4 (DPP4) inhibitors, in the whole population, without important differences by sex (28). We caution, however, that the statistical methods in that study were not as we would recommend.

Varadhan and colleagues highlight that the goal of confirmatory HTE analysis is to rigorously test hypotheses. Required elements include prespecification of subgroups, strong biological rationale and prior evidence to support subgroup hypotheses, adequate power to test subgroup hypotheses, prespecification of the analytic plan, control of family wise type I error, and the presence of a significant overall treatment effect (13). A distinguishing feature of confirmatory analysis is that uncertainty in the results, such as the CI surrounding the effect estimate, can be validly interpreted.

Effect estimates from observational studies are susceptible to confounding. In addition to stratification and model-based adjustment, investigators often implement propensity score methods. However, the propensity score generated in the whole population cannot be used to estimate subgroup-specific treatment effects (29, 30). Rather, one can estimate a propensity score within each prespecified subgroup for either matching or inverse probability weighting to achieve balance and control for confounding within the subgroups (31, 32). The size of subgroups should be sufficiently large to enable robust estimation of stratified propensity scores and matching. These methods to control for confounding cannot control for imbalance regarding unmeasured or coarsely measured covariates.

Confirming the presence or absence of HTE is also at the core of investigating the generalizability of trial evidence to subgroups poorly represented in the seminal RCTs (for example, ethnic minority members who are older with comorbid conditions). Such persons might be found in adequate numbers in RWD to permit subgroup analyses. In addition, replicating trial results in subgroups using RWD from persons *like* the trial participants can suggest that the observational methods are generating valid results in the other subgroups.

Objective 2: Describe the Magnitude and Nature of HTE

Descriptive HTE is the process of reporting on treatment effects, and their CIs, in prespecified subgroups but without testing hypotheses about the differences between subgroups or about differences between subgroup effects and the overall treatment effect (13). The primary objective—to estimate and report the magnitude of the treatment effect in known subgroups of interest—is valuable for later use in meta-analyses or for planning additional or larger studies. Investigators should report effect estimates within key subgroups with corresponding measures of precision. A forest plot is an acceptable first step for describing the consistency (or lack thereof) of treatment effect across important baseline characteristics, one variable at a time.

An illustration is found in a retrospective cohort study of management of severe carotid stenosis comparing carotid endarterectomy (CEA) to carotid artery stenting (CAS). The investigators report on the outcomes of stroke and death for the patients stratified by frailty to describe HTE in response to these treatments (33). For nonfrail patients, there was no important difference in the rate of the 30-day combined outcome between the CEA and CAS groups (CEA, 2.4% vs. CAS, 1.9%; $P = 0.59$). However, when compared with the CAS group, the CEA group had a higher rate of the outcome in prefrail patients (CEA, 2.9% vs. CAS, 1.0%; $P < 0.001$), frail patients (CEA, 3.9% vs. CAS, 1.2%; $P < 0.001$), and severely frail patients (CEA, 6.5% vs. CAS, 3.0%; $P = 0.04$).

Generally, the choice of subgroups for descriptive HTE analyses depends on the purpose of the study and should be prespecified in the study protocol. Potentially important classes of variables may be the intrinsic, extrinsic, and behavioral variables described in "What Are the Primary Sources of HTE?" as sources of HTE. Rothwell provided an extensive set of potential determinants of HTE (34). Descriptive HTE might be important for subpopulations for which limited evidence is available from trials, including subgroups defined by several variables, for example, older non-White women who are, in general, poorly represented in trials. Therefore, describing treatment effects using RWD for such underrepresented subgroups is valuable. Although investigators may start with binary explorations of subgroups (akin to what is reported in forest plots accompanying trial reports), they should also consider multivariable models that include higher level interactions.

Descriptive HTE Based on Bayesian Methods

A Bayesian subgroup analysis can be an effective approach for descriptive HTE when the subgroups are prespecified in the protocol (35). As described in the previous paragraph, the subgroups might be defined with several variables (for example, young White men, older Black women, and so forth). A Bayesian approach addresses the problem of large variance due to small subgroups by combining the subgroup treatment effects using a hierarchical model, where the subgroups are nested within the overall study. Estimated subgroup effects using Bayesian approaches will be a compromise between the subgroup-specific effect and the overall average effect. This type of subgroup estimation is known as shrinkage estimation, and the concept has a long history in statistical science and especially in the analysis of mixed effects models in which individual patient predictions are

important (36). The same concept can apply to subgroups of patients.

In the simplest case, where all subgroups are assumed to be similar in their characteristics, the degree of compromise depends on the variance of the treatment effect *within* the subgroup and the variance *between* the different subgroup treatment effects. If the *within* variance is large, then the treatment effect of that particular subgroup will be shifted close to the overall treatment effect, which is good, because a large *within* subgroup variance is a warning that the subgroup-specific effect is not trustworthy. A Bayesian approach does not borrow information indiscriminately, but borrows so that information on treatment effect is shared among persons with similar covariates. User-friendly software is now available for performing Bayesian subgroup shrinkage estimation (37).

Bayesian methodology for subgroup analysis, although designed for analysis of RCTs, applies to nonexperimental studies using RWD. Investigators need only unconfounded subgroup-specific effects, and their variances, to use the methodology. These subgroup estimates can be obtained from propensity score–based matching or by weighting within each subgroup, as was discussed in "Objective 1: Confirm Subgroup HTE" about confirmatory HTE analyses.

Descriptive HTE Based on Prognostic and Propensity Scores

There are 2 fundamental approaches to describing heterogeneity based on summary scores; 1 captures HTE dependent on the risk for outcome and the other captures HTE dependent on the probability that a person receives a treatment of interest. In the first approach, Kent and colleagues advocate for the use of the baseline risk for the outcome, a prognostic score, as the dimension along which to describe HTE. Baseline risk is the probability of the outcome in the absence of the treatment (9, 10). Treatment effect is then estimated within each quintile, for example, of the prognostic score. Although formal tests of heterogeneity may be conducted to assess whether the treatment effect varies across the quintiles, Kent and colleagues have generally emphasized a visual–qualitative approach, particularly for descriptive purposes (9). If available, an externally validated prognostic scoring system can be used, provided it calibrates well to the available data. For example, a safety study of the effect of tenofovir on kidney disease outcomes used a validated chronic kidney disease risk score (called Data Collection on Adverse Events of Anti-HIV Drugs chronic kidney disease risk) (38) and revealed little heterogeneity in the risk for the outcome among tenofovir-exposed and -unexposed persons when stratified by risk (Table 1). When a validated prognostic scoring system is not available, a scoring system can be developed (3). Modeling HTE as a function of a "treatment response score" is similar to use of a prognostic risk score but does not require separate estimation of the prognostic score (39).

The second approach to descriptive HTE uses the propensity score: the estimated probability of receiving treatment (compared with nontreatment) as a function of baseline covariates. The propensity score is used similarly to the prognostic score. For example, Kurth and colleagues used propensity score adjustment methods for confounding adjustment when evaluating the effect of tissue plasminogen activator on stroke mortality (40). Importantly, they also demonstrated an important gradient in the treatment effect across levels of the propensity score. Patients with high probabilities of being treated with tissue plasminogen activator were more likely to benefit than patients who were less likely to be treated; as shown, subgroups defined by the likelihood of being treated fared differently with treatment (Table 2).

Objective 3: Discover Subgroups With Important HTE

In contrast with objective 1, which focuses on hypothesis testing, this exploratory objective identifies subgroups that might benefit from treatment more than the average patient and with lower risk for harm. An important study objective might be to identify subgroups that *should* be further evaluated using rigorous observational designs with RWD or RCTs. Here, there is less concern about adequate power and multiple comparisons, but attention to bias and confounding cannot be ignored. For example, the oncology literature has reported exploratory studies of a tumor marker that is predictive of greater or lesser response to treatment. A recent example is the expression of the CD155 ligand on melanoma cells and on non–small cell lung cancer cells. In 2 different observational studies, people whose tumors expressed this marker had less response to anti-PD1 therapies than people whose tumors did not express these markers (41, 42). This discovery of HTE should prompt hypothesis-driven studies to confirm that this tumor marker is causally related to diminished benefit from these immunotherapies and to quantify the difference across groups to inform clinical decision making.

Ruberg and Shen describe the key elements to prespecify in subgroup search (43): 1) the method to be used

Table 1. Association Between TDF Exposure, D:A:D CKD Risk Strata, and Incidence of CKD*

TDF/D:A:D Risk Group	Unadjusted OR (95% CI)	Adjusted OR (95% CI)
No TDF		
Low-risk	1.00 (Ref)	1.00 (Ref)
Medium-risk	4.69 (1.70–12.96)	2.32 (0.72–7.52)
High-risk	37.56 (17.20–82.02)	19.55 (7.35–52.00)
TDF		
Low-risk	0.42 (0.16–1.11)	0.55 (0.19–1.54)
Medium-risk	5.37 (2.40–12.01)	3.96 (1.38–11.39)
High-risk	18.30 (8.42–39.78)	12.84 (4.57–36.07)

CKD = chronic kidney disease; D:A:D = Data Collection on Adverse Events of Anti-HIV Drugs; OR, odds ratio; TDF = tenofovir disoproxil fumarate.

* This table illustrates that the odds of the CKD outcome increase across the predicted risk strata and vary little by exposure: the incidence rate of the CKD outcome for the low-risk group with no TDF is 0.0017 versus 0.0006 with TDF (38).

(Table body reproduced from R Hsu, L Brunet, J Fusco, A Beyer, G Prajapati, C Wyatt, M Wohlfeiler, G Fusco, Risk of chronic kidney disease in people living with HIV by tenofovir disoproxil fumarate (TDF) use and baseline D:A:D chronic kidney disease risk score, *HIV Medicine* [John Wiley & Sons Ltd on behalf of British HIV Association], Volume 22, Issue 5, May 2021, Pages 325-333.)

Table 2. Proportion of Deaths Among 6,269 Ischemic Stroke Patients Registered in a German Stroke Registry Between 2000 and 2001 Who Were Treated or Not Treated With Tissue Plasminogen Activator, According to Percentiles of the Propensity Score for the Entire Study Population

Percentile	Treated (n = 212)				Not treated (n = 6,057)				Empirical OR*
	Score†	No.	Deaths		Score†	No.	Deaths		
			No.	%			No.	%.	
99 to 100	0.5809	36	3	8.3	0.5474	26	7	26.9	0.25
95 to <99	0.3143	73	13	17.8	0.2912	178	27	15.2	1.21
90 to <95	0.1393	55	8	14.6	0.1363	258	19	7.4	2.14
75 to <90	0.0585	31	3	9.7	0.0459	910	82	9.0	1.08
50 to <75	0.0115	10	4	40.0	0.0084	1,558	87	5.6	11.27
25 to <50	0.0017	5	2	40.0	0.0014	1,561	54	3.5	18.60
10 to <25	0.0004	2	1	50.0	0.000267	940	36	3.8	25.11
5 to <10		0	0	0	0.000066	313	6	1.9	
1 to <5		0	0	0	0.000027	251	8	3.2	
0 to <1		0	0	0	0.000007	62	1	1.6	
Overall	0.2521	212	34	16.0	0.0262	6057	327	5.4	3.35

*Propensity-stratum-specific-treatment–mortality odds ratio.
†Mean propensity score in percentile.
This table illustrates that the odds ratio (OR) varies inversely with the propensity of being treated, that is, individuals with a lower propensity score had higher odds of mortality (40).
(Table title and body reproduced from Tobias Kurth, Alexander M. Walker, Robert J. Glynn, K. Arnold Chan, J. Michael Gaziano, Klaus Berger, James M. Robins, Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect, *American Journal of Epidemiology* [Johns Hopkins Bloomberg School of Public Health in association with the Society for Epidemiologic Research], 2006, Volume 163, Issue 3, Pages 262-270, by permission of Oxford University Press.)

for exploration, 2) the list of potential predictive biomarkers (subgroup-defining variables), 3) how continuous predictors will be categorized to define subgroups, 4) other choices to be made in the analyses, 5) how adjustment for multiplicity will be done, and 6) bias correction, that is, how the estimated treatment effects in subgroups will be corrected for selection. Methods for subgroup discovery generally applicable to observational RWD include recursive partitioning (44–46), modified covariate regression (47), and a Bayesian method (48, 49).

Exploration for effect modification typically examines interactions between baseline characteristics of the person and the treatment. However, this approach is insufficient if the effect modifier is time-varying (such as the underlying disease severity) or if it is affected by prior treatment, as is often the case for chronic conditions where treatment decisions are updated on the patient's evolving condition. Often, both time-varying modifiers and time-varying treatments are present; conventional regression and propensity score stratification methods that naively condition on time-varying modifiers affected by prior treatment may yield biased estimates of modified treatment effects (50, 51). Methods such as g-estimation or structural nested mean models (52, 53) and history-adjusted marginal structural models might apply (54, 55).

Objective 4: Predict Individualized Treatment Effects

Like RCTs, many effectiveness studies using RWD are designed to test whether an intervention is efficacious *on average*. Ideally, we would like to estimate the treatment effect for an individual. However, this is not possible because each person receives only 1 of the comparison treatments. We can, however, estimate *individualized* treatment effects, which rely on modeling assumptions about how the treatment effect varies according to individual characteristics. Individualized estimates, also known as conditional average treatment effects (CATEs), may be more realistically estimated with RWD than in RCTs given the size and richness of RWD. El Sanadi and colleagues, as an example, used their institution's electronic medical record to create an online tool to assist the clinician and patient in selecting the next medication to add to metformin to improve outcomes associated with type 2 diabetes (56). They hypothesized that the effects of the treatment options would depend on characteristics of the individual patients. They developed parsimonious prediction models for each of 5 clinical outcomes of interest, following the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guidelines (Figure 2).

Estimation of CATE using machine-learning algorithms is an active area of statistical research. The proposed methods can be categorized into 2 large groups: meta-learners or modified machine-learning methods. Künzel and colleagues used a machine-learning algorithm that uses regression or classification as the "base learner" (for example, lasso, random forest, gradient boosting, or neural networks); they then estimated the expected outcome as a function of covariates in the treatment and comparison groups separately to assess for HTE (57). The modified machine-learning methods are mostly tree-based algorithms that seek to estimate the CATE directly, such as causal forest (58), causal boosting (59), and Bayesian regression tree models (60). Causal forest, for example, works by first finding a set of covariate values ("a neighborhood") where the treatment effect is constant but differs from the treatment effect in other neighborhoods; it then estimates the CATE over the neighborhoods.

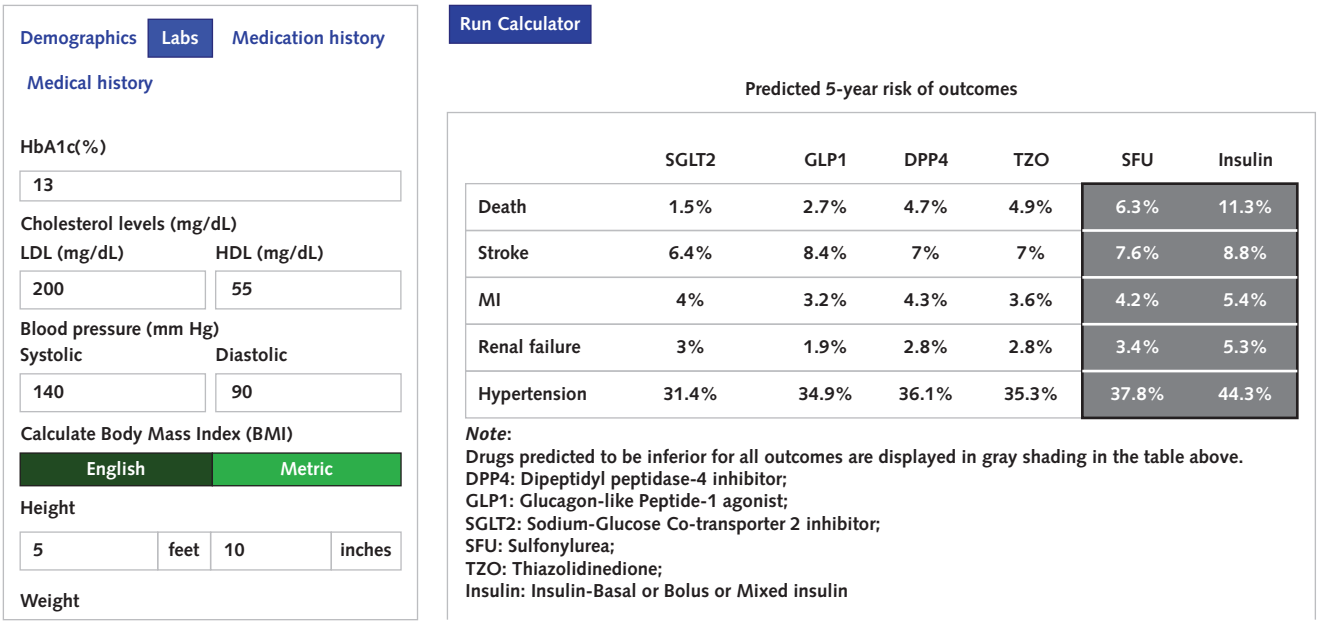
These methods are based in the frequentist framework. Henderson and colleagues developed an alternative, that is, a fully nonparametric, Bayesian machine-learning algorithm for individualized estimation of treatment effects in an RCT that might apply equally to RWD (61). This approach can quantify overall HTE, identify important patient characteristics related to HTE, estimate the proportion who benefit from the treatment, identify patient subpopulations deriving most benefit from treatment, detect crossover (qualitative) interactions, identify patients who are harmed by treatment, estimate individualized treatment effects, and predict treatment effect for a future patient. In reporting predictions for individuals, investigators should adhere to the TRIPOD Initiative guidelines for the reporting of studies developing, validating, or updating a prediction model (62). Although the models referred to in the TRIPOD guidelines are for the purposes of prognosis or diagnosis, they are also relevant to treatment effect prediction models.

What Is Still Needed?

Our work extends existing recommendations about HTE evaluation (for example, the PATH statement) in 2 main respects: 1) it is tailored to RWD rather than RCTs—the burgeoning literature on RWD does not address HTE assessment and 2) it addresses a broad range of study objectives from confirmatory to exploratory assessments. We distinguish here between 4 objectives that investigators may specify when investigating HTE: confirming subgroup HTE, describing the magnitude of HTE, discovering subgroups, and predicting individualized treatment effects. Here, we make recommendations for a principled approach to assessing HTE when using RWD that should complement those for HTE evaluation in the setting of RCTs (for example, FDA guidance on subgroup analysis [63], ICH-E9 [64], PATH statement [3]). Yet, more work is needed.

Outstanding issues include interpretation of results in subgroups when the biological plausibility or social rationale for differences in outcomes is not strong. In other words, there is a need for principled approaches to using information about newly discovered subgroups. Recommendations are needed for presenting results clearly to decision makers, including on what scale and with the best graphical approach for communicating HTE. Methods are urgently needed to address the problem of applicability of RCT evidence to subpopulations that are poorly represented in trials. Most importantly, we need to develop a framework for determining whether evidence on HTE is actionable for decision makers. For example, do clinicians make different prescribing decisions when presented with real-world HTE findings? Will payers limit coverage of products to subgroups that showed little benefit from the treatment? Should coverage with evidence development studies as requested by the Centers for Medicare &

Figure 2. Screenshot of the clinical decision support tool "labs" page and prediction outputs for an example "Patient X".



Patient X is an example of a standard patient whose demographics, laboratory values, and medical history were entered into the tool to provide a sample output. The image illustrates how individualized treatment effect estimates can be presented (56). BMI = body mass index; DPP4 = dipeptidyl peptidase-4 inhibitor; GLP1 = glucagon-like peptide-1 agonist; HbA1c = hemoglobin A1c; HDL = high-density lipoprotein; insulin = insulin-basal or bolus or mixed insulin; LDL = low-density lipoprotein; MI = myocardial infarction; SFU = sulfonylurea; SGLT2 = sodium–glucose cotransporter-2 inhibitor; TZO = thiazolidinedione. (Image reproduced from *Endocrine Practice* [American Association of Clinical Endocrinology], Volume 27, Caroline E. El Sanadi, Kevin M. Pantalone, Xinge Ji, Michael W. Kattan, Development and internal validation of a prediction tool to assist clinicians selecting second-line therapy following metformin monotherapy for type 2 diabetes, Pages 334-341, Copyright (2021), with permission from Elsevier.).

Medicaid Services require HTE evaluation, especially for very high-cost drugs (for example, chimeric antigen receptor T cells in oncology)?

Furthermore, there is a need for methods that can incorporate sources of heterogeneity beyond patient-level characteristics, including provider-level and health system–level factors. Hierarchical modeling approaches to incorporate these sources of HTE are needed, particularly techniques for quantifying the magnitude of their contributions to the overall HTE. Hierarchical propensity score modeling is one possibility (65). In the absence of peer-reviewed reporting guidelines on HTE in the context of RWD, an important next step will be to develop consensus on methods to evaluate HTE, followed by promulgation of expert-based guidelines.

We are optimistic that there will be increasing rigor in the use of RWD to generate reliable evidence to inform the care of patients. We urge clinicians and investigators to consider HTE always when interpreting the results of studies and when generating new evidence because HTE is almost certainly present and there are valid methods to study it.

From Johns Hopkins University School of Medicine, Baltimore, and Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland (J.B.S., R.V.); Leiden University Medical Center, Leiden, the Netherlands (R.H.H.G.); Department of Biostatistics, University of Michigan, Ann Arbor, Michigan (N.C.H.); Harvard Medical School Department of Population Medicine and Harvard Pilgrim Health Care Institute, Boston, Massachusetts (X.L.); Jikei University School of Medicine, Tokyo, Japan (K.N.); Teva Pharmaceutical Industries, Petah Tikva, Israel (S.K.); Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland (S.A., J.H.); School of Public Health and Community Medicine, Institute of Medicine, University of Gothenburg, Gothenburg, Sweden (F.N.); and Merck & Co., Rahway, New Jersey (M.B.).

Note: Dr. Jodi B. Segal, Associate Editor for *Annals*, had no role in the editorial review of or decision to publish this article.

Acknowledgment: This manuscript is endorsed by the International Society for Pharmacoepidemiology (ISPE).

Grant Support: Dr. Varadhan's work was supported through Regional Oncology Research Center grant P30CA006973 from the National Cancer Institute.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M22-1510.

Corresponding Author: Jodi B. Segal, MD, MPH, 624 North Broadway, Room 644, Baltimore, MD 21205; e-mail, jsegal@jhmi.edu.

Correction: This article was corrected on 25 April 2023 to add an author whose name had been omitted. A correction has been published (doi:10.7326/L23-0162)

Author contributions are available at Annals.org.

References

1. **Bica I, Alaa AM, Lambert C, et al.** From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. Clin Pharmacol Ther. 2021;109:87-100. [PMID: 32449163] doi:10.1002/cpt.1907
2. **U.S. Food and Drug Administration (FDA).** Framework for FDA's Real-World Evidence Program. FDA; December 2018.
3. **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. [PMID: 31711134] doi:10.7326/M18-3667
4. **Yusuf S, Wittes J, Probstfield J, et al.** Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA. 1991;266:93-8. [PMID: 2046134]
5. **Oxman AD, Guyatt GH.** A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116:78-84. [PMID: 1530753] doi:10.7326/0003-4819-116-1-78
6. **Wang R, Lagakos SW, Ware JH, et al.** Statistics in medicine—reporting of subgroup analyses in clinical trials. N Engl J Med. 2007;357:2189-94. [PMID: 18032770] doi:10.1056/NEJMsr077003
7. **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:E901-6. [PMID: 32778601] doi:10.1503/cmaj.200077
8. **Kent DM, Alsheikh-Ali A, Hayward RA.** Competing risk and heterogeneity of treatment effect in clinical trials [Editorial]. Trials. 2008;9:30. [PMID: 18498644] doi:10.1186/1745-6215-9-30
9. **Kent DM, Nelson J, Dahabreh IJ, et al.** Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol. 2016;45:2075-88. [PMID: 27375287] doi:10.1093/ije/dyw118
10. **Kent DM, Rothwell PM, Ioannidis JP, et al.** Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11:85. [PMID: 20704705] doi:10.1186/1745-6215-11-85
11. **Hernán MA, Robins JM.** Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183:758-64. [PMID: 26994063] doi:10.1093/aje/kwv254
12. **Hernán MA, Alonso A, Logan R, et al.** Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008;19:766-79. [PMID: 18854702] doi:10.1097/EDE.0b013e3181875e61
13. **Varadhan R, Segal JB, Boyd CM, et al.** A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2013;66:818-25. [PMID: 23651763] doi:10.1016/j.jclinepi.2013.02.009
14. **FUTURE II Study Group.** Quadrivalent vaccine against human papillomavirus to prevent high-grade cervical lesions. N Engl J Med. 2007;356:1915-27. [PMID: 17494925] doi:10.1056/NEJMoa061741
15. Food and Drug Administration Amendments Act of 2007, HR 3580, 110th Congress (2007-2008).
16. **Sutton AJ, Cooper NJ, Abrams KR, et al.** A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. J Clin Epidemiol. 2005;58:26-40. [PMID: 15649668]
17. **Vickers AJ, Van Calster B, Steyerberg EW.** Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. [PMID: 26810254] doi:10.1136/bmj.i6
18. **Henderson NC, Varadhan R.** Bayesian bivariate subgroup analysis for risk–benefit evaluation. Health Serv Outcomes Res Method. 2018;18:244-64. doi:10.1007/s10742-018-0188-1
19. **Berger ML, Sox H, Willke RJ, et al.** Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making.

Pharmacoepidemiol Drug Saf. 2017;26:1033-9. [PMID: 28913966] doi:10.1002/pds.4297

20. **Rudrapatna VA, Butte AJ.** Opportunities and challenges in using real-world data for health care. J Clin Invest. 2020;130:565-574. [PMID: 32011317] doi:10.1172/JCI129197

21. **Wang SV, Pinheiro S, Hua W, et al.** STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. BMJ. 2021;372:m4856. [PMID: 33436424] doi:10.1136/bmj.m4856

22. **Westreich D.** Epidemiology by Design: A Causal Approach to the Health Sciences. Oxford Univ Pr; 2020.

23. **Knol MJ, VanderWeele TJ.** Recoding preventive exposures to get valid measures of interaction on an additive scale [Letter]. Eur J Epidemiol. 2011;26:825-6. [PMID: 21892791] doi:10.1007/s10654-011-9613-2

24. **Knol MJ, VanderWeele TJ, Groenwold RH, et al.** Estimating measures of interaction on an additive scale for preventive exposures. Eur J Epidemiol. 2011;26:433-8. [PMID: 21344323] doi:10.1007/s10654-011-9554-9

25. **Lesko CR, Henderson NC, Varadhan R.** Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2018;100:22-31. [PMID: 29654822] doi:10.1016/j.jclinepi.2018.04.005

26. **von Elm E, Altman DG, Egger M, et al; STROBE Initiative.** The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147:573-7. [PMID: 17938396] doi:10.7326/0003-4819-147-8-200710160-00010

27. **Nyirjesy P, Sobel JD, Fung A, et al.** Genital mycotic infections with canagliflozin, a sodium glucose co-transporter 2 inhibitor, in patients with type 2 diabetes mellitus: a pooled analysis of clinical studies. Curr Med Res Opin. 2014;30:1109-19. [PMID: 24517339] doi:10.1185/03007995.2014.890925

28. **Lega IC, Bronskill SE, Campitelli MA, et al.** Sodium glucose cotransporter 2 inhibitors and risk of genital mycotic and urinary tract infection: a population-based study of older women and men with diabetes. Diabetes Obes Metab. 2019;21:2394-404. [PMID: 31264755] doi:10.1111/dom.13820

29. **Green KM, Stuart EA.** Examining moderation analyses in propensity score methods: application to depression and substance use. J Consult Clin Psychol. 2014;82:773-83. [PMID: 24731233] doi:10.1037/a0036515

30. **Rassen JA, Shelat AA, Myers J, et al.** One-to-many propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf. 2012;21 Suppl 2:69-80. [PMID: 22552982] doi:10.1002/pds.3263

31. **Dong J, Zhang JL, Zeng S, et al.** Subgroup balancing propensity score. Stat Methods Med Res. 2020;29:659-76. [PMID: 31456486] doi:10.1177/0962280219870836

32. **Yang S, Li F, Thomas LE, et al.** Covariate adjustment in subgroup analyses of randomized clinical trials: a propensity score approach. Clin Trials. 2021;18:570-81. [PMID: 34269087] doi:10.1177/17407745211028588

33. **Chan V, Rheaume AR, Chow MM.** Impact of frailty on 30-day death, stroke, or myocardial infarction in severe carotid stenosis: endarterectomy versus stenting. Clin Neurol Neurosurg. 2022;222:107469. [PMID: 36228442] doi:10.1016/j.clineuro.2022.107469

34. **Rothwell PM.** Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet. 2005;365:176-86. [PMID: 15639301] doi:10.1016/S0140-6736(05)17709-5

35. **Henderson NC, Louis TA, Wang C, et al.** Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. Health Serv Outcomes Res Methodol. 2016;16:213-233. [PMID: 27881932]

36. **Dixon RH, Laszlo J.** Utilization of clinical chemistry services by medical house staff. An analysis. Arch Intern Med. 1974;134:1064-7. [PMID: 4433187]

37. **Wang C, Louis TA, Henderson NC, et al.** beanz: an R package for Bayesian analysis of heterogeneous treatment effects with a graphical user interface. J Stat Soft. 2018;85:1-31.

38. **Hsu R, Brunet L, Fusco J, et al.** Risk of chronic kidney disease in people living with HIV by tenofovir disoproxil fumarate (TDF) use and baseline D:A:D chronic kidney disease risk score. HIV Med. 2021;22:325-33. [PMID: 33247876] doi:10.1111/hiv.13019

39. **Kovalchik SA, Varadhan R, Weiss CO.** Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Stat Med. 2013;32:4906-23. [PMID: 23788362] doi:10.1002/sim.5881

40. **Kurth T, Walker AM, Glynn RJ, et al.** Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol. 2006;163:262-70. [PMID: 16371515] doi:10.1093/aje/kwj047

41. **Jiang C, Qu X, Ma L, et al.** CD155 expression impairs anti-PD1 therapy response in non-small cell lung cancer. Clin Exp Immunol. 2022;208:220-32. [PMID: 35262683] doi:10.1093/cei/uxac020

42. **Lepletier A, Madore J, O'Donnell JS, et al.** Tumor CD155 expression is associated with resistance to anti-PD1 immunotherapy in metastatic melanoma. Clin Cancer Res. 2020;26:3671-81. [PMID: 32345648] doi:10.1158/1078-0432.CCR-19-3925

43. **Ruberg SJ, Shen L.** Personalized medicine: four perspectives of tailored medicine. Stat Biopharm Res. 2015;7:214-29. doi:10.1080/19466315.2015.1059354

44. **Foster JC, Taylor JM, Ruberg SJ.** Subgroup identification from randomized clinical trial data. Stat Med. 2011;30:2867-80. [PMID: 21815180] doi:10.1002/sim.4322

45. **Lipkovich I, Dmitrienko A, Denne J, et al.** Subgroup identification based on differential effect search–a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011;30:2601-21. [PMID: 21786278] doi:10.1002/sim.4289

46. **Su X, Tsai C-L, Wang H, et al.** Subgroup analysis via recursive partitioning. J Mach Learn Res. 2009;10:141-58.

47. **Tian L, Alizadeh AA, Gentles AJ, et al.** A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2014;109:1517-32. [PMID: 25729117] doi:10.1080/01621459.2014.951443

48. **Berger JO, Wang X, Shen L.** A Bayesian approach to subgroup identification. J Biopharm Stat. 2014;24:110-29. [PMID: 24392981] doi:10.1080/10543406.2013.856026

49. **Ruberg SJ.** Assessing and communicating heterogeneity of treatment effects for patient subpopulations: the hardest problem there is. Pharm Stat. 2021;20:939-44. [PMID: 33655601] doi:10.1002/pst.2110

50. **Robins J.** A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Dis. 1987;40 Suppl 2:139S-61S. [PMID: 3667861] doi:10.1016/s0021-9681(87)80018-8

51. **Robins J.** The control of confounding by intermediate variables. Stat Med. 1989;8:679-701. [PMID: 2749074] doi:10.1002/sim.4780080608

52. **Robins JM.** Association, causation, and marginal structural models. Synthese. 1999;121:151-79.

53. **Robins JM.** Structural nested failure time models. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics: Survival Analysis. J Wiley; 1998:4372-89.

54. **Petersen ML, Deeks SG, Martin JN, et al.** History-adjusted marginal structural models for estimating time-varying effect modification. Am J Epidemiol. 2007;166:985-93. [PMID: 17875580] doi:10.1093/aje/kwm232

55. **van der Laan MJ, Petersen ML.** History-adjusted marginal structural models and statically-optimal dynamic treatment regimes. U.C. Berkeley

Division of Biostatistics Working Paper Series. University of California, Berkeley; 2004. Working Paper 158.
56. **El Sanadi CE, Pantalone KM, Ji X, et al.** Development and internal validation of a prediction tool to assist clinicians selecting second-line therapy following metformin monotherapy for type 2 diabetes. Endocr Pract. 2021;27:334-41. [PMID: 33685669] doi:10.1016/j.eprac.2020.10.015
57. **Künzel SR, Sekhon JS, Bickel PJ, et al.** Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A. 2019;116:4156-65. [PMID: 30770453] doi:10.1073/pnas.1804597116
58. **Athey S, Wager S.** Estimating treatment effects with causal forests: an application. Observational Studies. 2019;5:37-51. doi:10.1353/obs.2019.0001
59. **Powers S, Qian J, Jung K, et al.** Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med. 2018;37:1767-1787. [PMID: 29508417] doi:10.1002/sim.7623
60. **Hahn PR, Murray JS, Carvalho CM.** Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Analysis. 2020;15:965-1056. doi:10.1214/19-BA1195
61. **Henderson NC, Louis TA, Rosner GL, et al.** Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. Biostatistics. 2020;21:50-68. [PMID: 30052809] doi:10.1093/biostatistics/kxy028
62. **Collins GS, Reitsma JB, Altman DG, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63. [PMID: 25560714] doi:10.7326/M14-0697
63. **U.S. Food and Drug Administration (FDA).** Collection, Analysis, and Availability of Demographic Subgroup Data for FDA-Approved Medical Products. FDA; 2013.
64. **International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).** E9 Statistical Principles for Clinical Trials - Scientific Guideline. ICH Harmonised Tripartite Guideline. European Medicines Agency; 1998.
65. **Li F, Zaslavsky AM, Landrum MB.** Propensity score weighting with multilevel data. Stat Med. 2013;32:3373-87. [PMID: 23526267] doi:10.1002/sim.5786

Author Contributions: Conception and design: J.B. Segal, R. Varadhan, R.H.H. Groenwold, X. Li, K. Nomura, S. Kaplan, F. Nyberg, M. Burcu.
Analysis and interpretation of the data: J.B. Segal, R. Varadhan, M. Burcu, N.C. Henderson.
Drafting of the article: J.B. Segal, R. Varadhan, R.H.H. Groenwold, X. Li, K. Nomura, F. Nyberg, M. Burcu.
Critical revision of the article for important intellectual content: J.B. Segal, R. Varadhan, R.H.H. Groenwold, X. Li, S. Kaplan, F. Nyberg, M. Burcu.
Final approval of the article: J.B. Segal, R. Varadhan, R.H.H. Groenwold, X. Li, K. Nomura, S. Kaplan, S. Ardeshirrouhanifard, J. Heyward, F. Nyberg, M. Burcu, N.C. Henderson.
Statistical expertise: R. Varadhan, R.H.H. Groenwold, X. Li, N.C. Henderson.
Administrative, technical, or logistic support: J.B. Segal, J. Heyward.
Collection and assembly of data: J.B. Segal, S. Ardeshirrouhanifard, J. Heyward, N.C. Henderson.

Appendix Table. Glossary of Terms

Terms	Description
Overall treatment effect	A comparison of response between 2 groups that comprise the entire study sample, where each group is exposed to a different treatment.
Bayesian inference versus frequentist inference	Frequentist inference is a statistical framework that evaluates the population parameters by imagining repeated samples from an appropriate model. The population parameters are assumed to be fixed, but unknown. Bayesian inference is a framework that uses prior beliefs or information and updates those beliefs based on the observed data to derive probabilistic statements about unknown population parameters, using an appropriate model for the data-generating process. Here, the population parameters are random and unknown. Both frequentist and Bayesian frameworks require a data-generating model, but the Bayesian framework also requires a prior distribution for population parameters. In the frequentist framework, the parameters are fixed but the data are random, whereas in the Bayesian framework, the data are fixed and the parameters are random.
Conditional average treatment effect (CATE)	A model-based estimate of the individual treatment effect where a model depicting the relationship between the outcome, treatment, and covariates is fitted. Then, CATE is calculated for each individual in a study sample as a contrast of their model-estimated response under 2 treatments.
Effect modification	A measure of how the treatment effect varies according to different values of a covariate. Effect modification is commonly assessed by including a treatment by covariate product term in a regression model. For example, the coefficient of age–treatment product term is a measure of how the treatment effect varies as age varies.
Effectiveness	The performance of an intervention in the setting in which it is usually used in practice.
Efficacy	The performance of an intervention under ideal and controlled circumstances.
Heterogeneity of treatment effect (HTE)	The explainable (nonrandom) variation in treatment response that can be attributed to differences in patient characteristics.
Individual treatment effect	A comparison of an individual's response under 2 different treatments. This is often unobservable because any individual can only be exposed to 1 treatment (unless the condition being treated is acute).
Individualized treatment effect	See conditional average treatment effect.
Interaction	Same as effect modification in terms of statistical description, but quite different conceptually. Interaction is said to exist between 2 manipulable variables, whereas effect modification measures how 1 manipulable variable varies as a function of a fixed covariate. Interaction can be synergistic or antagonistic.
Posterior distribution	A probability distribution that reflects the researcher's belief about a population parameter of interest after observing the data.
Prior distribution	A probability distribution that reflects the researcher's belief about a population parameter of interest before observing the data.
Qualitative HTE	A variation in treatment effect, of the opposite direction, according to levels of covariate. For example, men have a beneficial effect from the treatment, but women have a harmful effect.
Quantitative HTE	A variation in treatment effect, of the same direction, according to levels of a covariate. For example, men and women both have a beneficial effect from the treatment, but the magnitude of benefit is significantly different.
Real-world data (RWD)	Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.
Real-world evidence	Clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.
Shrinkage estimation	Treatment effect in a subgroup is estimated as a compromise between the "raw" or "observed" treatment effect in that group and the overall (average) treatment effect. The degree of compromise depends on the size of the subgroup and the shrinkage method. The smaller the subgroup the greater the compromise.
Subgroup analysis	The most popular way of examining HTE, in which the entire study sample is divided into mutually exclusive groups and the treatment effect is estimated in each group—for example, the treatment effect in men and in women.
Generalizability/ Transportability	Pertains to whether the evidence on benefits and risks of an intervention obtained from a controlled clinical trial is valid when applied to patients in the real world.
Applicability	Pertains to whether the evidence on benefits and risks of an intervention obtained from a controlled clinical trial is relevant and valid for a particular subpopulation of at-risk individuals. The distinction between applicability and generalizability is that applicability requires that we define a specific subpopulation, for example, Hispanic, women, older than 70 years, with diabetes.
Treatment effect scale	The scale in which treatment effect is measured. For example, this could be a ratio of average response under treatment to the average response without treatment (relative scale), or it could be the difference in average response under treatment to the average response without treatment (absolute scale).