SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

*Original Research Article*

Statistical Methods in Medical Research
2025, Vol. 34(4) 648–662

DOI: 10.1177/09622802251316969
journals.sagepub.com/home/smm
Sage

# Generalized framework for identifying meaningful heterogenous treatment effects in observational studies: A parametric data-adaptive G-computation approach

**Roch A. Nianogo[1,2], Stephen O'Neill[3] and Kosuke Inoue[4,5]**

**Abstract**

There has been a renewed interest in identifying heterogenous treatment effects (HTEs) to guide personalized medicine. The objective was to illustrate the use of a step-by-step transparent parametric data-adaptive approach (the generalized HTE approach) based on the G-computation algorithm to detect heterogenous subgroups and estimate meaningful conditional average treatment effects (CATE). The following seven steps implement the generalized HTE approach: Step 1: Select variables that satisfy the backdoor criterion and potential effect modifiers; Step 2: Specify a flexible saturated model including potential confounders and effect modifiers; Step 3: Apply a selection method to reduce overfitting; Step 4: Predict potential outcomes under treatment and no treatment; Step 5: Contrast the potential outcomes for each individual; Step 6: Fit cluster modeling to identify potential effect modifiers; Step 7: Estimate subgroup CATEs. We illustrated the use of this approach using simulated and real data. Our generalized HTE approach successfully identified HTEs and subgroups defined by all effect modifiers using simulated and real data. Our study illustrates that it is feasible to use a step-by-step parametric and transparent data-adaptive approach to detect effect modifiers and identify meaningful HTEs in an observational setting. This approach should be more appealing to epidemiologists interested in explanation.

**Keywords**

Heterogeneity, effect modification, clustering, machine learning, regularization, causal inference, explanation, epidemiology

## 1 Introduction

In public health and clinical settings, it is not uncommon to observe some heterogeneity in intervention or treatment effects—which occurs when the effect of an intervention or treatment is different from one subgroup to another.[1–3] Epidemiologists have been long interested in investigating heterogenous treatment effects (HTEs) for several reasons.[1,2,4] First, it is well-known that one-size-fits-all treatments across individuals could lead to disparate benefits and harms in some groups. For instance, statins are not recommended for all patients but only for those with a certain risk factor profile—this is to maximize benefit among those at risk and limit the harms in subgroups less at risk.[5] Second, tailoring interventions or policies and targeting subgroups who are most at risk or who would benefit the most given limited resources could potentially lead to the greatest societal impact.[6–9] This is particularly relevant in disparities research as the detection of differences

[1] Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles (UCLA), USA
[2] California Center for Population Research, University of California, Los Angeles (UCLA), USA
[3] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, UK
[4] Department of Social Epidemiology, Graduate School of Medicine, Kyoto University, Japan
[5] Hakubi Center, Kyoto University, Japan

**Corresponding author:**
Roch A. Nianogo, Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles (UCLA), Los Angeles, CA 90095, USA.
Email: niaroch@ucla.edu

in treatment effects across socio-economic or racial/ethnic subgroups could help identify potential barriers and levers for achieving optimal health for all, as failing to identify such barriers could lead to unavoidable deaths and exacerbate existing disparities.[10–15] More recently, there has been an increased interest in identifying HTE to guide personalized medicine[16–18] as well as to further investigate negative trials, that is, those that report a null total effect, as it could reveal the presence of treatment benefit for some subgroups (e.g. those whose body mass index is $\geq 30\,\text{kg/m}^2$) but not all.[19–21]

## 1.1 Brief summary of conventional versus modern approaches to detecting heterogeneity

The identification of HTEs, also known as effect modification or moderation is not a novel concept in epidemiology,[3,22] however, its recent application in personalized interventions has given researchers a renewed appreciation for its importance. Conventional methods have long made use of statistical interactions and parametric models to detect heterogeneity.[3,22] Over the past decade, many tools designed to detect HTEs have emerged across disciplines such as computer science, statistics, econometrics, and epidemiology. These innovations have yielded "non-parametric" estimators and have introduced techniques to detect HTEs leveraging data-adaptive (machine learning) algorithms. Many of these approaches for estimating individual-level effects belong to the family of 'meta-learner' approaches[23] in that they either: (a) estimate a single model for all units [S-learners], (b) estimate separate models for treated and control units [T-learners], (c) model the propensity for treatment as well as modelling outcomes [X-learners] or (d) 'orthogonalize' outcomes and treatment assignment prior to estimation (that is, they use residuals from models in place of observed variables) [R-learners]. For example, causal forest, a commonly used R-learner approach[24–28] uses an ensemble of trees optimized to detect HTEs, allowing us to estimate Conditional Average Treatment Effect (CATE) for each individual as a function of multi-dimensional individual characteristics. Causal forests differ from random forests in that they predict the treatment effect (e.g. outcome improvement due to the treatment) for each individual and seek to identify its variations, whereas random forests primarily focus on predicting the outcome itself. To reduce the risk of overfitting, causal forests also employ an approach, called "honest estimation" when building each decision tree, ensuring an observation is not simultaneously used to determine sample splits and to estimate effects.

## 1.2 Three challenges when considering machine learning or conventional approaches to detect HTEs and estimate subgroup effects

When opting for novel machine learning (e.g. causal forest) or conventional approaches (e.g. parametric models with interaction terms) to detect HTEs and estimate subgroup effects, three potential challenges are at the forefront of the epidemiologist's mind: (a) the problem of accuracy versus interpretability, (b) the problem of underfitting or overfitting when using parametric models, and (c) the problem of type 1 error and joint testing.

### *1.2.1 The problem of accuracy versus interpretability*

Many non-parametric machine-learning algorithms such as neural networks and random forests are excellent in producing accurate predictions; however, this often comes at the cost of model interpretability, since they are highly non-linear (neural networks) or average across many decision trees (random forests), making them somewhat of a 'black box' where the relationship between inputs and outputted predictions is opaque. This has led to a growing literature on interpretability and explainable machine-learning modeling.[29] On the other hand, traditional parametric linear or logistic models while highly interpretable can produce poor prediction accuracy. This phenomenon known as the "two cultures" is hardly a novel problem and has been well described by Breiman (2001) who lamented the underuse of machine learning over conventional parametric modeling for solving prediction problems.[30] With more recent advances in both "cultures", it is conceivable to envision a scenario where both approaches/cultures are used to produce models with good enough prediction accuracy and acceptable interpretability. One such example is that of regularization methods (e.g. ridge regression) which are class of "interpretable"[31] parametric machine-learning algorithms that do not sacrifice too much interpretability but with much-improved performance accuracy.

Furthermore, selecting an adequate modeling approach is particularly relevant when one is interested in an explanation/causal inference problem as opposed to a purely predictive problem per se.[31] In fact, knowledge about the underlying data-generating process or causal structural model is crucial in guiding the inclusion of variables in one's model when the goal is to elucidate the causal relationship between exposure and outcome.[32] The investigator would include variables sufficient for controlling confounding in their model while excluding descendants of the primary exposure of interest, such as colliders.[33–35]

### *1.2.2 The problem of underfitting or overfitting when using parametric modeling*

Underfitting occurs when a model is not sufficiently rich to capture the true relationship between covariates and outcomes, while overfitting occurs when the model is too flexible capturing idiosyncratic variation and leading to poor out-of-sample performance. Thus, there is a bias variance trade-off when specifying models. Data-driven machine-learning approaches assess model performance out of sample allowing these risks to be balanced. It has been recognized that parametric regressions can achieve similar estimates to non-parametric, data-driven approaches, when models are flexibly specified allowing for rich interactions between covariates and treatment[3] but at the cost of overfitting.[36]

For prediction tasks, one could fit a saturated parametric model with reasonable n-way interaction terms and then use a shrinkage-based method to correct for the presence of unnecessary variables such as by shrinking their coefficient toward zero. Such methods include regularization methods like Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO),[37] and Elastic Net models.[38] The use of regularization methods has been shown to be superior to the ordinary least square (OLS) model when the regularization hyperparameters are well-tuned such as through cross-validation.[38]

Where interest centers on estimating effects in observational settings, a post-double selection approach,[39] where variables that are selected as being predictive of outcomes or treatment assignment are chosen for inclusion can also be used to reduce bias. For subgroup analysis, consideration of variables predictive of treatment-covariate interactions may also be necessary.[40]

Additional approaches can include the use of inverse probability of treatment weights (IPTW) for adjusting for residual confounding.[41]

### *1.2.3 The problem of type 1 error and joint testing*[42,43]

After estimating individual-level treatment effects, one is often interested in aggregating these for subgroups of interest. The most-principled approach is arguably to pre-specify the groups for which effects will be presented,[44] while data-adaptive approaches to identify subgroups have also been proposed.[45–51] A number of studies investigating HTEs when using machine-learning methods report the effect of the exposure on the outcome across subgroups defined by combinations of covariates and oftentimes further report such effects across all covariates in the model.[21,36] Reporting the effects across all covariates and then highlighting the ones that are “significant” can be seen at best as a hypothesis-generating endeavor rather than confirmatory[52] and at worst as a form of significance chasing. The latter can lead to false positives[53] and threats of type 1 error,[54,55] particularly when multiple testing adjustments (e.g. Bonferroni correction, Semi-Bayes,[56–58] Benjamini-Hochberg) are not accounted for.

In particular, our proposed approach deals with type 1 error in the following manner. First, it employs regularization methods such as LASSO[37] or Elastic Net models[38] which effectively performs variable selection by shrinking less important coefficients to zero. This reduces the number of hypotheses tested and controls Type I error by penalizing the inclusion of non-informative variables. Second, cross-validation helps in assessing the model’s performance and increases its generalization capability. This reduces the risk of Type I errors by ensuring that the model’s performance is not overly optimistic due to overfitting. Third, achieving balance in covariates between candidate subgroups can help prevent the false discovery of HTE subgroups as demonstrated by Rigdon et al.[42] Fourth, our approach helps reduce the risk of type I error by only evaluating effects across the subgroups and variables that have been (pre)-identified as potential effect modifiers by our generalized HTE approach.

In the current paper, we describe the use of a theory- and data-driven approach to detect meaningful HTEs (*the generalized HTE approach*) and estimate subgroup treatment effects using the G-computation algorithm,[59] while recognizing these 3 challenges.

**Box.** Definitions

1. **Effect modifier/moderator**: a variable that differentially modifies the observed effect of the intervention on the outcome.
2. **Average Treatment Effect (ATE)**: the expected effect of an intervention in the total population, that is $\tau = E(Y^1 - Y^0)$
3. **Heterogenous Treatment Effects (HTE)**: difference in effects of an intervention across individuals or groups.
4. **Conditional Average Treatment Effects (CATEs)**: the expected effect for individuals with a particular covariate profile (e.g. $W = w$), that is, a contrast of the *expected* potential outcomes for that individual under each treatment: $\tau(w) = E(Y^1 - Y^0 | W = w)$. It is also sometimes referred to as individualized treatment effect (ITE) when each individual is defined by the values of $W$.

5. **Subgroup Average Treatment Effects**: the average effect for individuals in a particular group (e.g. $G = g$), that is $\tau(g) = E(Y^1 - Y^0|G = g)$

## 2 Methods

In implementing the *generalized HTE approach*, we make use of a parametric "interpretable" supervised machine-learning algorithm using regularization methods such as LASSO and a split sample approach (i.e. using the training subsample for model development and test subsample for making model prediction) to prevent overfitting and improve the generalization performance of the model. We also use an unsupervised machine-learning algorithm such as clustering to identify effect modifiers. We then obtain effect estimates using G-computation (also referred to as recycled predictions[60,61] or direct standardization) by contrasting estimated potential outcomes under each treatment. G-computation is a generalization of the standardization method for time-varying settings that has been used in the epidemiologic literature to (a) obtain standardized estimate over all covariates and interactions,[62–64] (b) adjust for time-varying confounders,[65–68] and (c) to project the impact of hypothetical interventions.[69–72]

### 2.1 Notations and definitions

Let $Y$ be the outcome, $X$ the exposure or treatment, $Y^{X=x}$ the potential outcome $Y$ under treatment $X = x$, $C$ a set of confounders sufficient for confounding control and $W$ a set of potential effect modifiers.

### 2.2 Identifiability conditions

We assume conditional exchangeability (unconfoundedness or ignorability) given confounders, $C$ within $W = w$. In other words, the potential outcome $Y^X$ had treatment $X$ been set to $x$, is independent of treatment $X$ given adjusted confounding variables $C$ within $W = w$; where $C$ satisfies the backdoor criterion within $W = w$,[73,74] that is, contains no descendants of $X$ (e.g. colliders) and that upon conditioning on $C$, there is no open backdoor path between $X$ and $Y$ within $W = w$. Importantly, the confounder set $C$ includes all potential modifiers, $W$ that are also confounders (i.e. when $C$ and $W$ are not mutually exclusive and there exist covariates that are in both $C$ and $W$). Other assumptions include positivity (common support), consistency (treatment irrelevance),[75] no interference or SUTVA (stable unit treatment value assumption),[76] no model functional form misspecifications, no measurement error and no selection bias.

### 2.3 Estimation of conditional average treatment effects

When the aforementioned identifiability conditions are satisfied, we can estimate the CATE using observational data. More formally, CATE can be derived from the following equation.

$$\begin{aligned}\text{CATE} &= \tau(w) = E(Y^{X=1} - Y^{X=0}|W = w)\\ &= E(Y^{X=1}|\ W = w) - E(Y^{X=0}|\ W = w)\\ &= \sum_{C=c} E(Y|X = 1,\ W = w, C = c)\ P(C = c|W = w)\\ &\quad - \sum_{C=c} E(Y|X = 0,\ W = w, C = c)\ P(C = c|W = w)\end{aligned}$$

To date, several non-parametric algorithms such as meta-learners including causal forests have been used to estimate CATE. Nevertheless, the factors influencing the estimated effects obtained using these non-parametric approaches may not always be as transparent or readily interpretable as those from conventional parametric models. Therefore, in the subsequent section, we delineate each step of a parametric approach, *the generalized HTE approach*, to detect meaningful HTEs and estimate CATE.

### 2.4 G-computation steps in identifying HTE and estimating CATE

#### 2.4.1 Step 1: select variables that satisfy the backdoor criterion and potential effect modifiers

In this essential step, the investigator identifies the set of covariates sufficient for confounding control that satisfy the backdoor criterion,[74] and selects potential effect modifiers with particular care taken to remove potential colliders (to avoid collider-stratification bias[33]). This can be aided by the use of directed acyclic diagrams[34] and background knowledge[32] which can be particularly important when using data-driven approaches for variable selection.[77]

### 2.4.2 Step 2: specify a flexible saturated model including potential confounders and potential effect modifiers

An appropriate parametric model should be specified (e.g. Gaussian, Binomial). To capture heterogeneity, the investigator could specify a saturated model, that is a model that includes all possible n-way interaction terms between the treatment indicator and other included variables.

### 2.4.3 Step 3: apply a selection method to reduce overfitting

This step can occur simultaneously with the previous step. Here, regularization methods such as Ridge regression, LASSO or elastic net could be used to reduce overfitting.

LASSO aims to find the set of coefficients that minimizes the sum-of-squared errors subject to a constraint on the sum of absolute values of coefficients. A penalty function is added to the typical Ordinary Least Squares (OLS) loss function, as follows:

$$\text{Loss} = \sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{J}\hat{\beta}_j c_{ij}\right)^2 + \lambda \sum_{j=1}^{J}|\hat{\beta}_j|$$

where $Y$ represents the dependent variable, $c_{ij}$ is the $j$th of $J$ (rescaled) covariates for individual $i$ and $\beta_j$ is the corresponding coefficient. The tuning parameter, $\lambda$, determines the extent to which the model complexity is penalized, with larger values resulting in more variables being excluded.

In contrast, and for illustration purposes, Ridge regression, shrinks coefficients toward 0, with smaller coefficients shrunk more but unlike LASSO, does not exclude variables.

$$\text{Loss} = \sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{J}\hat{\beta}_j c_{ij}\right)^2 + \lambda \sum_{j=1}^{J}\hat{\beta}_j^2$$

For Elastic Net, the loss function to be minimized is:

$$\text{Loss} = \sum_{i=1}^{N}\left(Y_i - \sum_{j=1}^{J}\hat{\beta}_j c_{ij}\right)^2 + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{J}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{J}|\hat{\beta}_j|\right)$$

Note that when $\alpha = 1$, elastic net is equivalent to LASSO, while when $\alpha \to 0$, elastic net approaches ridge regression. When $\lambda = 0$, the penalty function corresponds to the typical OLS loss function.

In effect, Elastic Net models are types of more transparent parametric supervised machine-learning algorithms that simultaneously allow for model selection, the shrinking of some coefficients toward zero (e.g. Ridge regression) and even the setting of some coefficients to zero (e.g. LASSO). This allows for a sparse and parsimonious model and avoids overfitting. In fact, the goal of these regularization methods is to avoid overfitting by adding some penalization for each additional term in the model. This step requires, however, the tuning of hyperparameters (lambda and alpha), typically done through cross-validation.

While the above approaches can be readily applied when interest is on avoiding overfitting in predictions, when interest is in estimating causal effect, these approaches may not adequately control for potential confounders in some cases (e.g. when confounders are selected out of the model). Including additional variable selection steps that identify variables influencing both treatment decisions and outcome and aided by background knowledge,[32] can be beneficial. Re-estimating the outcome model using the union of variables selected can mitigate this concern and provide valid inference.[39] This is also relevant when the interest is in effect modification as performing variable selection on interactions between modifiers and treatment is also important and recommended,[40] albeit it may increase the risk of selecting bad controls,[77] so care must be taken to consider the appropriateness of selected covariates. The additional variable selection step is known as the post-double selection[39] and is done on the training sample.

Additionally, as interests are in causal effects, one can instead use IPTW to adjust for residual confounding,[41] and achieve balance.

### 2.4.4 Step 4: predict potential outcomes under treatment and no treatment

Using the fitted model developed in the training data (with the hyperparameters selected in the model development in step 3), we can then predict potential outcomes under treatment, $\hat{Y}_i^{X=1}$, and under no treatment, $\hat{Y}_i^{X=0}$, in the testing/validation data.

### 2.4.5 Step 5: obtain a contrast between potential outcomes for each individual

The contrast between the potential outcomes under the different treatment scenarios can be expressed for instance as mean difference (MD; a difference between means) or risk differences (RDs; difference between predicted probabilities). Such individual-level CATE estimates are conditioned on each value of individual's characteristics and are sometimes called individualized treatment effects (ITE).

### 2.4.6 Step 6: fit cluster modeling to find appropriate clusters of effects and identify potential effect modifiers

Once individual-level CATE estimates are obtained, they can be aggregated to obtain subgroup effects. This endeavor can be aided by using clustering algorithms, which are a type of unsupervised machine-learning technique that are helpful in finding subgroups of similar individuals. This machine-learning procedure avoids having to find groups manually, reducing the risk of data-mining. It systematically and iteratively uses an algorithm to calculate the distance between each formed cluster. It then aims at reducing the distance within a subgroup and increasing the distance between groups. The k-means cluster algorithm, for instance, aims to partition $N$ observations into $k$ clusters/groups $(g_1, \ldots , g_k)$ such that the squared Euclidean distance between the row vector for any individual's variables (effect modifiers here) $(m_i)$ and the centroid vector of their respective cluster is at least as small as the distance to the centroids of the remaining clusters[78] that is, they are assigned to the cluster they are most similar to in terms of the variables in $m$. An iterative procedure is used to determine the centroids and the cluster membership.[78] Another similar clustering approach include the Ward hierarchical clustering[79] which minimizes the total within-cluster variance aiming at minimizing the Euclidian distance.

Clustering can be used to identify potential effect modifiers. This is done using the following sub-steps. First, once individual CATEs have been estimated, a clustering algorithm can be used to identify clusters (typically but not always, two clusters are identified) in which individuals differ as much as possible between clusters (in terms of estimated effects here). These clusters are akin to the leaf nodes of a decision tree. Of note, some algorithms such as $k$-means or Ward clustering need to have a large number of distinct values that is greater than the number of clusters to be searched. Here, we can use a statistical trick by adding some small random noise (called jittering) to each CATE to help identify clusters. Jittering has been used in machine-learning model training and has been shown to help with model generalization.[80,81] Note that this trick is only for identifying the best/optimal number of clusters. Second, each covariate $z$-scores across the identified clusters are estimated. Third, the variable importance calculated as the difference in covariate $z$-scores across clusters is also estimated. Selected potential effect modifiers are those whose variable importance exceeds a certain threshold, for instance, we use 0.2 in our illustration.

### 2.4.7 Step 7: estimate subgroup CATEs in a marginal structural model

To estimate subgroup CATEs, we can include interaction terms between effect modifier(s) identified in Step 5 and the treatment in an OLS regression using the entire dataset. The subgroup CATEs are then simply obtained by finding the treatment effect across the subgroup defined by the identified effect modifiers in a marginal structural model (e.g. using G-computation or IPTW). If IPTW is used, robust sandwich estimators are used to estimate standard errors and 95% confidence intervals (CIs) and if G-computation, then bootstrap is used instead. Other estimators could also be used including propensity score matching (e.g. Ridge matching)[82] doubly robust estimation,[83] and other semiparametric estimators of the ATE (*Busso M, DiNardo J, McCrary J. Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects, 2009, Unpublished*).

## 3 Illustration

As a proof of concept, we conducted two studies. First, we conduct a simulation study to assess the ability of *the generalized HTE approach* to detect HTEs and estimate CATEs. Second, we implement the method in an applied study using the Health and Retirement Study (HRS)[84] to identify potential effect modifiers of the association between baseline low kidney function (as measured by an estimated glomerular filtration, eGFR) and incident dementia.

We next describe the implementation of the methods for both studies. We first fit a LASSO model on the training data (a random 70% of the data) to select relevant variables and interaction terms. The tuning parameters (lambda) were obtained via 10-fold cross-validation. We used inverse probability of treatment weighting (IPTW) to adjust for potential confounding. The fitted model was used on the testing data (the remaining 30%) to predict the potential outcomes under treatment and no treatment. We then took the contrast in terms of the difference between the potential outcomes to obtain the CATE estimates.

The Ward clustering approach[79] which minimizes the total within-cluster variance was used to identify distinct clusters based on the CATE estimates. This approach minimized the Euclidian distance and selected the optimal number of clusters based on the Beale index. We identified potential effect modifiers by examining the variable importance calculated as

**Table 1.** Subgroup CATE on the Additive and Relative Scale Presenting the Effect of Intensive Blood Pressure Control Treatment on Risk of a Cardiovascular Disease Event Across Moderators Identified Via the Generalized HTE Approach in the Simulated Data A1, $N = 10{,}000$.

| Variables and threshold | *N* | Proportion (%) | CATE risk difference (RD, 95%CI) | *P for interaction (RD)* | CATE risk ratio (RR, 95%CI) | *P for interaction (RR)* |
|---|---|---|---|---|---|---|
| Overall | 100 | 10,000 | −0.03 (−0.05, −0.02) | – | 0.82 (0.75, 0.91) | – |
| eGFR < 73 | 51.59 | 5159 | −0.12 (−0.14, −0.09) | 0 | 0.51 (0.44, 0.59) | 0 |
| eGFR ≥ 73 | 48.41 | 4841 | 0.05 (0.03, 0.08) | | 1.37 (1.20, 1.56) | |
| Aspirin = 0 | 48.83 | 4883 | −0.21 (−0.23, −0.19) | 0 | 0.25 (0.21, 0.30) | 0 |
| Aspirin = 1 | 51.17 | 5117 | 0.13 (0.11, 0.16) | | 2.19 (1.93, 2.50) | |

eGFR: estimated glomerular filtration rate; RD: risk difference, RR: risk ratio; CI: confidence interval.

the difference in covariate $z$-scores across clusters. In fact, selected potential effect modifiers were those whose variable importance was greater than 0.2. After identifying the potential effect modifiers, we fit a marginal structural model with robust standard error to estimate CATEs in the subgroups defined by the potential effect modifiers.

## 3.1 Application to simulated data

### 3.1.1 Overview

In the simulation study, we employed a data-generating process in which we simulated an observational study inspired from the randomized controlled trial simulation study described in Rigdon et al.[42] (see eSection 1 and eSection 2 for details on the data-generating mechanism and implementation of generalized HTE approach). We simulated 10,000 individuals (simulation A1) of whom about 30% were taking blood pressure medication to achieve a systolic blood pressure <120 mmHg to reduce cardiovascular outcomes[85] (i.e., intensive blood pressure control) (see eTable 1 for baseline characteristics). To assess the method's capabilities, we also simulated different sizes (simulation A2, $N = 100{,}000$ and simulation A3, $N = 1000$). These three simulation studies used the same data-generating process. We report the results of simulation A1 in the main manuscript and appendix (Table 1, eFigure 1, eFigure 2, eFigure 3, eTable 2) and those of the simulation A2 and simulation A3 in the appendix only (eFigure 4 and eFigure 5).

## 3.2 Results

### 3.2.1 Distribution of CATEs

Plotting the distribution of the CATEs (eFigure 1), we can see that the distribution is not unimodal but rather bimodal. In simulation A1, one of the distributions is centered toward a negative effect while the other around a positive effect. This preliminary step helps evaluate the potential for heterogeneity before identifying the presence of effect modifiers. This bi-model distribution can also be seen in a plot of CATE estimates by rank of CATE estimates (eFigure 2).

### 3.2.2 Identifying potential effect modifiers

The variables whose importance exceeded a certain threshold, for example 0.2 were selected as potential effect modifier (eFigure 3).

In simulation A1 ($N = 10{,}000$), we identified two potential effect modifiers: aspirin use (a binary variable) and estimated glomerular filtration rate (eGFR, a continuous variable). Of note, the cut-off for binarizing eGFR was calculated as the average of the mean of eGFR across the two identified clusters. In our case, it was $(75.31 + 70.74)/2 \approx 73.0$ (eTable 2). The other variables had a variable importance <0.2. Examining the variable importance plot (eFigure 3) suggests that the observed heterogeneity in eFigure 1 could potentially be due to the presence of effect modifiers.

### 3.2.3 Estimating subgroup CATEs

In simulation A1, using the *generalized HTE approach*, we identified four meaningful subgroups defined by aspirin use or eGFR (Table 1): participants whose eGFR < 73 (RD: −0.12, 95%CI (−0.14, −0.09), participants whose eGFR ≥ 73 (RD: 0.05, 95%CI: 0.03, 0.08), participants not taking aspirin (RD: −0.21, 95%CI: −0.23, −0.19) and participants taking aspirin (RD: 0.13, 95%CI: 0.11, 0.16).

**Table 2.** Baseline Participants' Characteristics in the Health and Retirement Student (HRS), $N = 11{,}033$.

| Characteristics | Overall, $N = 11{,}033$ | Low eGFR, $N = 6519$ | Normal eGFR, $N = 4514$ |
|---|---|---|---|
| Older age (Age ≥ 65), n (%) | 6962 (63%) | 6155 (94%) | 807 (18%) |
| Female sex, n (%) | 6500 (59%) | 3766 (58%) | 2734 (61%) |
| Participant years of education, Mean (SD) | 12.7 (3.0) | 12.5 (3.1) | 13.0 (3.0) |
| Married, n (%) | 7034 (64%) | 3914 (60%) | 3120 (69%) |
| Non-NH White participants, n (%) | 2488 (23%) | 1269 (19%) | 1219 (27%) |
| Income < 130 of the FPL, n (%) | 1329 (12%) | 740 (11%) | 589 (13%) |
| Low child SES, n (%) | 5236 (47%) | 3646 (56%) | 1590 (35%) |
| APOE 4, n (%) | 2930 (27%) | 1690 (26%) | 1240 (27%) |
| Systolic Blood Pressure (mmHg), Mean (SD) | 131 (20) | 134 (21) | 128 (19) |
| HbA1c (%): NHANES Equivalent, Mean (SD) | 5.87 (0.98) | 5.88 (0.89) | 5.86 (1.11) |
| HDL (mg/dL); NHANES Equivalent, Mean (SD) | 54 (16) | 54 (16) | 55 (16) |
| BMI, kg/m$^2$, Mean (SD) | 29.4 (5.7) | 28.6 (5.4) | 30.4 (6.0) |
| Waist Circumference (cm), Mean (SD) | 40 (6) | 39 (6) | 40 (6) |
| C-reactive Protein (mg/L); NHANES Equivalent, Mean (SD) | 0.45 (0.87) | 0.44 (0.92) | 0.48 (0.78) |
| Pulse (Beats per Minute), Mean (SD) | 70 (11) | 69 (11) | 72 (11) |
| Exercise (MET ≥ 35), n (%) | 2598 (24%) | 1401 (21%) | 1197 (27%) |
| Never smoked, n (%) | 6266 (57%) | 3724 (57%) | 2542 (56%) |
| Number of alcoholic drinks in last 3 months, Mean (SD) | 0.69 (1.32) | 0.58 (1.14) | 0.85 (1.53) |
| CESD/Depression Score, Mean (SD) | 1.39 (1.94) | 1.28 (1.80) | 1.55 (2.11) |
| Anxiety Index, Mean (SD) | 1.57 (0.58) | 1.55 (0.56) | 1.60 (0.61) |
| State Anger Index, Mean (SD) | 1.50 (0.50) | 1.46 (0.48) | 1.55 (0.53) |
| Trait Anger Index, Mean (SD) | 2.18 (0.68) | 2.13 (0.66) | 2.26 (0.70) |
| Incident dementia, *n* (%) | 1651 (15%) | 1351 (21%) | 300 (6.6%) |

SD: Standard deviation; NH: not-Hispanic; SES: socio-economic status; APOE4: apolipoprotein E4 carrier; HbA1c: hemoglobin A1c; HDL: high-density lipoprotein; NHANES: National Health and Nutrition Examination Survey; MET: metabolic equivalent of task; CESD: Center for Epidemiologic Studies Depression Scale; FPL: federal poverty level; eGFR: estimated glomerular filtration rate. More details on variable definition can be found in the appendix (eSection 3).

## 3.3 Application to real data

### *3.3.1 Overview*

For this application, we use 2006–2020 data from the US Health and Retirement Study (HRS), a longitudinal panel study that surveys a representative sample of about 20,000 Americans aged >50 years and their spouses. More information can be found here.[84] In particular, we used the biomarker subsample data—a random 50% of the HRS data (in 2006 and 2008) which was preselected for the enhanced face-to-face interviews (EFTFI) which included collection of biomarker specimens (e.g. HbA1c). To obtain our analytical sample, we combined the 2006-subcohort with the 2008-subcohort; excluded individuals aged <50y and those with prevalent dementia at baseline in 2006/2008; implemented multiple imputation approach and used the first imputed dataset with complete data ($N = 11{,}033$, see Table 2 for baseline characteristics). It is known that low kidney function is positively associated with dementia.[86] In the current illustration, the objective was to identify potential effect modifiers of the association between baseline low kidney function (as measured by an estimated glomerular filtration, eGFR) and incident dementia. More details on variable definition can be found in the appendix (eSection 3).

## 3.4 Result

### *3.4.1 Distribution of CATEs*

Plotting the distribution of the CATEs (Figure 1), we can see that the distribution is not unimodal but rather bimodal or made of at least two distinct distributions with different dispersion and central tendency parameters. One of the distributions is centered toward zero while the other is centered toward a larger effect. This mixed distribution can also be appreciated in a plot of CATE estimates by rank of CATE estimates (eFigure 6).

### *3.4.2 Identifying potential effect modifiers*

As mentioned, the variables whose importance exceeded a certain threshold, for example 0.2 were included as potential effect modifier (Figure 2). In the HRS data, we identified three potential effect modifiers: older age (a binary variable, age

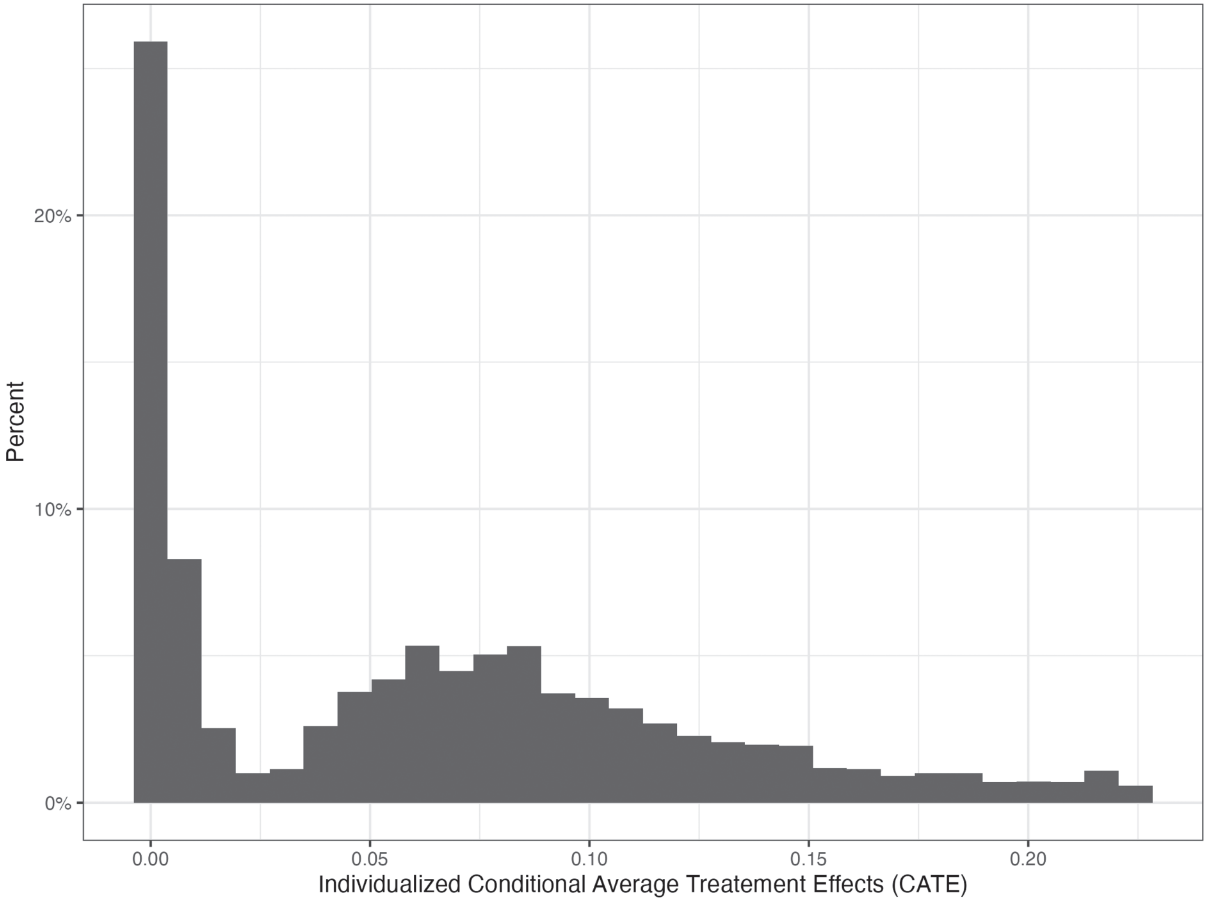

**Figure 1.** Histogram of the conditional average treatment effect (CATE) estimated via the generalized HTE approach in the health and retirement study ($N = 11{,}033$).

≥65 vs. <65), low childhood SES (a binary variable: low vs. high) and years of education (a continuous variable). For the continuous variable, the cut-off identified through the generalized HTE approach was calculated as the average of the years of education across the two identified clusters. In our case, it was $(13.15 + 11.96)/2 \approx 12.6$ (Table 3).

### *3.4.3 Estimating subgroup CATEs*

In the HRS dataset, using the *generalized HTE approach*, we identified six meaningful subgroups defined by older age, years of education or childhood SES (Table 4): Younger adult (RD: 0.00, 95%CI: [−0.02, 0.03]); older adults (RD: 0.10, 95%CI: [0.08, 0.13]) and p for interaction < 0.001; years of education < 12.6 (Risk Ratio [RR]: 1.59, 95%CI: [1.29, 1.95]); years of education ≥ 12.6 (RR: 2.43, 95%CI: [1.61, 3.65]) and p for interaction < 0.07. Moderate to high childhood SES (RD: 0.04, 95%CI: [0.01, 0.06]); low childhood SES (RD: 0.10, 95%CI: [0.07, 0.13]) and p for interaction < 0.001. Of note, older age was an effect modifier on both absolute and relative scale; years of education was an effect modifier on the relative scale only and childhood SES was an effect modifier on the additive scale only.

### *3.4.4 Shapley values*

We estimated the Shapley values by fitting the estimated conditional treatment effect (CATE) on covariates using a random forest model. In particular, the Shapley[87] values for a feature represent the average marginal contribution of that feature, for a specific observation/individual, to the prediction of CATE; this contribution is obtained by averaging across all possible combinations of features for that specific observation/individual. For instance, for observation 1, being less than 65 contributes the most negatively to the average CATE and low childhood SES contributes the most positively to the average CATE (see eFigure 7). Based on the mean absolute error, the most important feature across all individuals was the feature “older age” (see eFigure 8). The Shapley values were estimated with the help of the R package *iml*. We presented the feature value contribution and the feature importance for the first observation based on the estimated Shapley values in the

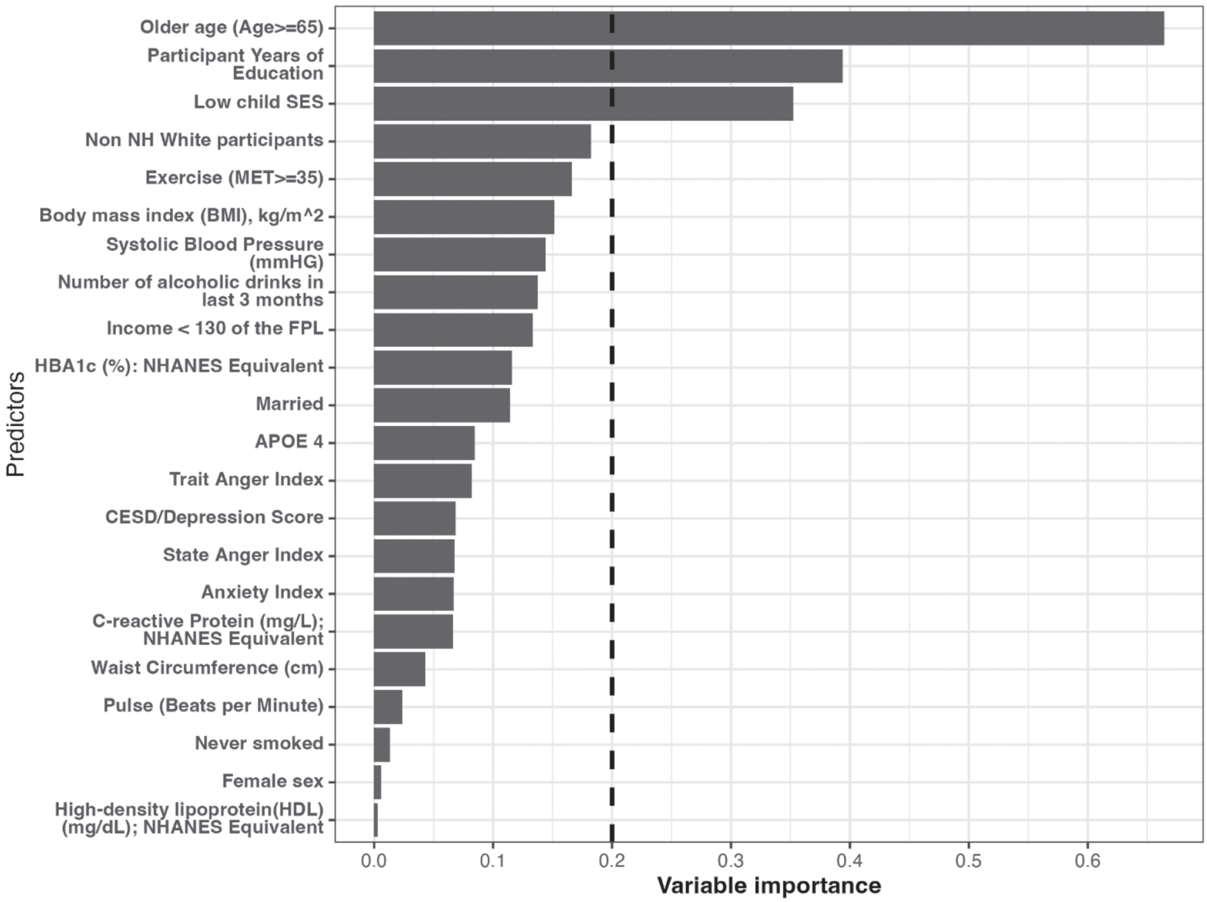

**Figure 2.** Variable importance estimated via the generalized HTE approach in the health and retirement study ($N = 11{,}033$). Only the top important covariates whose importance equal or exceed 0.20 were considered potential effect modifiers for the effect of treatment on the outcome.

**Table 3.** Cluster Effects and Effect Modifiers Identified Via the Generalized HTE Approach in the Health and Retirement Student (HRS), $N = 11{,}033$.

| Clusters | CATE (RD scale) | Older adult (≥ 65) (proportion) | Years of education (mean) | Low childhood SES (proportion) |
|---|---|---|---|---|
| 1 | 0.04 | 0.51 | 13.15 | 0.41 |
| 2 | 0.10 | 0.83 | 11.96 | 0.59 |

CATE: Individualized conditional average treatment effect; RD: risk difference.
The effect of low kidney function on dementia risk is higher among older adults, those with less years of education and who are more likely to come from a low-income household during childhood.

supplemental materials (eFigure 7, eFigure 8). The top 3 features (using the feature importance) were also those identified as potential effect modifiers using our generalized HTE approach.

# 4 Discussion

## 4.1 Summary and explanation of the main findings

The purpose of this paper was to illustrate, as a proof of concept, that we can use a step-by-step transparent parametric data-adaptive approach (the generalized HTE approach) based on the G-computation algorithm to detect heterogenous subgroups and identify meaningful HTEs. Using real data and simulated data, we were able to identify meaningful HTEs and estimate CATEs by implementing our 7-step data-adaptive approach.

**Table 4.** Subgroup CATE on the Additive and Relative Scale Presenting the Effect of Low Kidney Function on Incident Dementia Across Effect Modifiers Identified via the Generalized HTE Approach in the Health and Retirement Student (HRS), $N = 11,033$.

| Variables and threshold | *N* | Proportion (%) | CATE Risk difference (RD, 95%CI) | *P for interaction (RD)* | CATE Risk ratio (RR, 95%CI) | *P for interaction (RR)* |
|---|---|---|---|---|---|---|
| Total effect | 11,033 | 100 | 0.07 (0.05, 0.09) | – | 1.75 (1.45, 2.11) | – |
| Younger adult[a] (<65) | 4071 | 36.90 | 0.00 (−0.02, 0.03) | 0.00 | 1.07 (0.68, 1.68) | 0.02 |
| Older adult[a] (≥65) | 6962 | 63.10 | 0.10 (0.08, 0.13) | | 1.93 (1.55, 2.40) | |
| Years of education < 12.6 | 6071 | 55.03 | 0.08 (0.05, 0.11) | 0.43 | 1.59 (1.29, 1.95) | 0.07 |
| Years of education ≥ 12.6 | 4962 | 44.97 | 0.06 (0.04, 0.08) | | 2.43 (1.61, 3.65) | |
| Moderate to high childhood SES | 5797 | 52.54 | 0.04 (0.01, 0.06) | 0.00 | 1.52 (1.15, 2.02) | 0.26 |
| Low childhood SES | 5236 | 47.46 | 0.10 (0.07, 0.13) | | 1.89 (1.47, 2.42) | |

CI: confidence interval; SES: socio-economic status.

[a]Age was dichotomized to avoid violation of the positivity assumption.

The proposed approach was designed to address three particular challenges when investigating HTEs in epidemiology: (1) the problem of accuracy versus interpretability, (2) the problem of underfitting/overfitting when using parametric modeling, and (3) the problem of type 1 error and joint testing. To solve problems (1) and (2), a regularized transparent supervised machine-learning model such as LASSO, was chosen. The selection of variables to be included in the model was guided by both theory: causal diagrams and background knowledge[32] as well as by data-adaptive approaches: LASSO.[39] Lastly, to address problem (3) several strategies were used. First, it employs LASSO which performs variable selection reducing the number of hypotheses tested. Second, cross-validation helps in assessing the model's performance and increases its generalization capability and ultimately reducing the risk of Type I errors. Third, achieving balance in covariates (achieved in step 3 and 7, see eFigure 9 for covariate balance) between candidate subgroups can help prevent the false discovery of HTE subgroups as demonstrated by Rigdon et al.[42] Fourth, the subgroup effects estimated were only those estimated across identified potential effect modifiers (not across all available covariates as often done in the literature[21,36])

Our *generalized HTE approach* should in principle be more appealing to epidemiologists interested in explanation and causal inference questions given that: i) it is parametric and transparent, ii) it makes use of statistical interactions to detect heterogeneity, and iii) it explicitly strives to include variables that satisfy the backdoor criterion in an observational setting.

Furthermore, our approach and other methods purporting to identify HTE require sufficient amounts of data to have adequate power to detect heterogeneity. In addition, it may be necessary to evaluate heterogeneity on different scales, that is additive and multiplicative (which, our approach can naturally accommodate). In other words, when HTEs are truly present and there is no bias, failure to detect true HTEs could reflect lack of power or the scale chosen to evaluate heterogeneity. For instance, in our applied example, one variable (years of education) did not modify the effect of low kidney function on dementia risk on the absolute scale; however, it modified such effect on the relative scale. Furthermore, larger sample sizes relative to the number of variables are better for identifying true effect modifiers. In our simulation A3 ($N = 1000$), our approach identified in addition to the two known effect modifiers other variables for which there was little effect modification (eFigure 5). On the other hand, in the larger simulations A1 ($N$ = 10,000) and A2 ($N$ = 100,000), our approach correctly identified the two known effect modifiers (eFigure 3 and eFigure 4).

## 4.2 Limitations of the current approach, extension and future directions

The approach presented here is a proof of concept and as such does not aim to compare its performance to other HTE-based methods such as causal forest. Future studies should explore the impact of choosing different thresholds for identifying potential effect modifiers as well as the performance of the current approach with several model specifications (e.g. ridge regression), functional forms (e.g. Gaussian, binomial) and different clustering approaches (e.g. k-means, centroid, Ward) against other machine-learning methods including but not limited to causal forest and Bayesian causal forest. The comparison should assess metrics such as root mean square errors/accuracy, bias and coverage in detecting the correct subgroups.

## 5 Conclusion

We illustrated, as a proof of concept, a step-by-step transparent parametric data-adaptive approach (the generalized HTE approach) based on the G-computation algorithm to detect heterogenous subgroups and identify meaningful HTEs. Our generalized approach was able to identify meaningful HTEs and estimate CATE defined by effect modifiers. Future studies should compare the performance of this approach to other HTE methods in terms of bias, coverage and ability to detect subgroups. Our *generalized HTE approach* should in principle be more appealing to epidemiologists interested in explanation/causal inference questions given that it is parametric and transparent, it makes use of statistical interactions to detect heterogeneity, and explicitly strives to include variables that satisfy the backdoor criterion in an observational setting.

### Author contributions

RAN conceptualized the study. RAN, SON and KI contributed to the problem definition, conducted the data analysis, interpreted the results and wrote the first draft. All authors provided critical input and insights into the development and writing of the article and approved the final manuscript as submitted.

### Consent to participate

For the Health and Retirement Study (HRS) data, participants were provided with a written informed consent information document prior to each interview and then gave oral consent by agreeing to do the interview at the start of each interview. We used the collected publicly available de-identified data. In addition, consent to participate was not applicable for the simulation study.

### Data availability

The Health and Retirement Study (HRS) is a publicly available dataset that can be obtained through the University of Michigan's HRS data repository (https://hrsdata.isr.umich.edu/data-products/public-survey-data). In addition, the description of how to generate the simulated data is included in the supplemental material. The code to generate the data and to implement the generalized HTE approach can be found at https://github.com/nianogo/generalized_hte.

### Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Ethics approval

This study used publicly available de-identified data and simulated data and as such did not qualify as human subject research and ethics approval was not needed.

### Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: RN was partly supported by the National Institute on Minority Health and Health Disparities (grant number 1K01MD014163-01A1).

### ORCID iD

Roch A. Nianogo https://orcid.org/0000-0001-5932-6169

### Supplemental material

Supplemental material for this article is available online.

### References

1. Knol MJ and VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol* 2012; **41**: 514–520.
2. Van Der Weele TJ and Knol MJ. A tutorial on interaction. *Epidemiol Methods* 2014; **3**: 33–72.
3. Rothman K, Greenland S and Lash T. *Modern epidemiology*, 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008.
4. Rothman KJ, Greenland S and Lash TL. *Modern epidemiology*, 3rd ed. Philadelphia, PA: Wolters Kluwer Health Adis (ESP), 2011.
5. Blum CB, Eckel RH, Goldberg AC, et al. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults. *Am Heart Assoc Task Force Practice Guidel Circ* 2014; **129**: 1–45.
6. Frohlich KL and Potvin L. Transcending the known in public health practice: the inequality paradox: the population approach and vulnerable populations. *Am J Public Health* 2008; **98**: 216–221.
7. Lalonde M. A new perspective on the health of Canadians. *Vasa* 1974; **32**: 76.
8. Rose G. Sick individuals and sick populations. *Int J Epidemiol* 1985; **14**: 32–38.
9. Inoue K, Seeman TE, Horwich T, et al. Heterogeneity in the association between the presence of coronary artery calcium and cardiovascular events: a machine learning approach in the MESA study. *Circulation* 2023; **147**: 132–141. Epub ahead of print 31 October 2022.

10. Adler NE and Newman K. Socioeconomic disparities in health: pathways and policies. *Health Aff* 2002; **21**: 60–76.
11. Adler NE and Rehkopf DH. U.S. disparities in health: descriptions, causes, and mechanisms. *Annu Rev Public Health* 2008; **29**: 235–252.
12. Braveman P. Health disparities and health equity: concepts and measurement. *Annu Rev Public Health* 2006; **27**: 167–194.
13. Braveman P, Cubbin C, Marchi K, et al. Measuring socioeconomic status/position in studies of racial/ethnic disparities: maternal and infant health. *Public Health Rep* 2001; **116**: 449–463.
14. Hays SL, Riley P and Radley D,C, et al. Reducing Racial/Ethnic Disparities in Access to Care, https://www.commonwealthfund.org/publications/issue-briefs/2017/aug/reducing-racial-and-ethnic-disparities-access-care-has (2017, accessed 18 October 2018).
15. Oakes JM and Naimi AI. Mediation, interaction, interference for social epidemiology. *Int J Epidemiol* 2016; **45**: 1912–1914.
16. Basu S, Sussman JB and Hayward RA. Detecting heterogeneous treatment effects to guide personalized blood pressure treatment: a modeling study of randomized clinical trials. *Ann Intern Med* 2017; **166**: 354–360.
17. Inoue K, Athey S and Tsugawa Y. Machine-learning-based high-benefit approach versus conventional high-risk approach in blood pressure management. *Int J Epidemiol* 2023; **52**: 1243–1256. Epub ahead of print 2 August 2023.
18. Kent DM, Steyerberg E and Van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *Br Med J* 2018; **363**: k4245. Epub ahead of print 10 December 2018.
19. Basu S, Raghavan S, Wexler DJ, et al. Characteristics associated with decreased or increased mortality risk from glycemic therapy among patients with type 2 diabetes and high cardiovascular risk: machine learning analysis of the ACCORD trial. *Diabetes Care* 2018; **41**: 604–612.
20. Baum A, Scarpa J, Bruzelius E, et al. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. *Lancet Diabetes Endocrinol* 2017; **5**: 808–815.
21. Wiemken TL, Furmanek SP, Carrico RM, et al. Effectiveness of oseltamivir treatment on clinical failure in hospitalized patients with lower respiratory tract infection. *BMC Infect Dis* 2021; **21**: 1–7.
22. VanderWeele TJ. *Explanation in causal inference: methods for mediation and interaction*. New York: Oxford University Press, 2015.
23. 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**: 4156–4165.
24. Athey S and Wager S. Estimating treatment effects with causal forests: an application. *Obs Stud* 2019; **5**: 37–51.
25. Athey S, Tibshirani J and Wager S. Generalized random forests. https://doi.org/101214/18-AOS1709 2019; **47**: 1148–1178.
26. Athey S and Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci U S A* 2016; **113**: 7353–7360.
27. Hahn PR, Murray JS and Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *SSRN Electronic Journal* 2017; **15**: 965–1056.
28. Wager S and Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc* 2018; **113**: 1228–1242.
29. Linardatos P, Papastefanopoulos V and Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy* 2021; **23**: 18.
30. Breiman L. *Statistical modeling: the two cultures* (with comments and a rejoinder by the author). https://doi.org/101214/ss/1009213726 2001; **16**: 199–231.
31. Shmueli G. *To explain or to predict?* https://doi.org/101214/10-STS330 2010; **25**: 289–310.
32. Hernán MA, Hernández-Díaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation. *Am J Epidemiol* 2002; **155**: 176–184.
33. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol* 2010; **39**: 417–420.
34. Greenland S, Pearl J and Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**: 37–48.
35. Pearl J. *Causality: models, reasoning, and inference, second edition*. Cambridge: Cambridge University Press, 2011.
36. 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**: 923–936.
37. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodological)* 1996; **58**: 267–288.
38. James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. New York, NY: Springer, 2013.
39. Belloni A, Chernozhukov V and Hansen C. Inference on treatment effects after selection among high-dimensional controls. *Rev Econ Stud* 2014; **81**: 608–650.
40. Blackwell M and Olson MP. Reducing model misspecification and bias in the estimation of interactions. *Polit Anal* 2022; **30**: 495–514.
41. Robins JM, Hernán MA and Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.
42. Rigdon J, Baiocchi M and Basu S. Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials* 2018; **19**: 1–15.
43. Watson JA and Holmes CC. Machine learning analysis plans for randomised controlled trials: detecting treatment effect heterogeneity with strict control of type I error. *Trials* 2020; **21**: 1–10.
44. Sun X, Briel M, Walter SD, et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *Br Med J* 2010; **340**: 850–854.
45. Foster JC, Taylor JMG and Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011; **30**: 2867–2880.

46. 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**: 2601–2621.
47. Dwivedi R, Tan YS, Park B, et al. Stable discovery of interpretable subgroups via calibration in causal studies. *Int Stat Rev* 2020; **88**: S135–S178.
48. Loh WY, He X and Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Stat Med* 2015; **34**: 1818.
49. Su X, Tsai C-L, Wang H, et al. Subgroup analysis via recursive partitioning Chih-Ling Tsai. *J Mach Learn Res* 2009; **10**: 141–158.
50. Dusseldorp E and Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Stat Med* 2014; **33**: 219–237.
51. Hapfelmeier A, Ulm K and Haller B. Subgroup identification by recursive segmentation. *J Appl Stat* 2018; **45**: 2864–2887.
52. Ranstam J. Hypothesis-generating and confirmatory studies, Bonferroni correction, and pre-specification of trial endpoints. *Acta Orthop* 2019; **90**: 297.
53. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine–reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007; **357**: 2189–2194.
54. Gelman A, Hill J and Yajima M. Methodological studies why we (usually) don't have to worry about multiple comparisons. *J Res Educ Eff* 2012; **5**: 189–211.
55. Jean Dunn O. Multiple comparisons among means. *J Am Stat Assoc* 1961; **56**: 52–64.
56. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med* 1993; **12**: 717–736.
57. Witte JS, Greenland S, Haile RW, et al. Hierarchical regression analysis applied to a study of multiple dietary exposures and breast cancer. *Epidemiology* 1994; **5**: 612–621.
58. Greenland S. Hierarchical regression for epidemiologic analyses of multiple exposures. *Environ Health Perspect* 1994; **102**: 33–39.
59. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model* 1986; **7**: 1393–1512.
60. Basu A and Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* 2005; **6**: 93–109.
61. StataCorp. Stata Statistical Software: Release 18, https://www.stata.com/support/faqs/resources/citing-software-documentation-faqs/ (2023, accessed 3 December 2023).
62. Mokhayeri Y, Hashemi-Nazari SS, Khodakarim S, et al. Effects of hypothetical interventions on ischemic stroke using parametric G-formula. *Stroke* 2019; **50**: 3286–3288.
63. Naimi AI, Cole SR and Kennedy EH. An introduction to g methods. *Int J Epidemiol* 2017; **46**: dyw323.
64. Snowden JM, Rose S and Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol* 2011; **173**: 731–738.
65. Daniel RM, De Stavola BL and Cousens SN. Gformula: estimating causal effects in the presence of time-varying confounding or mediation using the G-computation formula. *Stata J* 2011; **11**: 479–517.
66. Daniel R, De Stavola B, Cousens S, et al. Causal mediation analysis with multiple causally-ordered mediators. *Biometrics.* 2015; **71**: 1–14.
67. Daniel RM, De Stavola BL, Cousens SN, et al. Causal mediation analysis with multiple mediators. *Biometrics* 2015; **71**: 1–14.
68. Daniel RM, Cousens SN, De Stavola BL, et al. Methods for dealing with time-dependent confounding. *Stat Med* 2013; **32**: 1584–1618.
69. Danaei G, Pan A, Hu FB, et al. Hypothetical midlife interventions in women and risk of type 2 diabetes. *Epidemiology* 2013; **24**: 122–128.
70. Nianogo RA, Wang MC, Wang A, et al. Projecting the impact of hypothetical early life interventions on adiposity in children living in low-income households. *Pediatr Obes* 2017; **12**: 398–405.
71. Rojas-Saunero LP, Hilal S, Murray EJ, et al. Hypothetical blood-pressure-lowering interventions and risk of stroke and dementia. *Eur J Epidemiol* 2021; **36**: 69–79.
72. Taubman SL, Robins JM, Mittleman MA, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol* 2009; **38**: 1599–1611.
73. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol* 2019; **34**: 211–219.
74. Vanderweele TJ and Shpitser I. A new criterion for confounder selection. *Biometrics* 2011; **67**: 1406–1413.
75. Cole SR and Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology* 2009; **20**: 3–5.
76. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc* 1980; **75**: 591–593.
77. Hünermund P, Louw B and Caspi I. Double machine learning and automated confounder selection: a cautionary tale. *J Causal Inference* 2023; **11**: 20220078.
78. Steinley D. K-means clustering: a half-century synthesis. *Br J Math Stat Psychol* 2006; **59**: 1–34.
79. Murtagh F and Legendre P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J Classif* 2014; **31**: 274–295.

80. Czarnecki WM and Podolak IT. Machine learning with known input data uncertainty measure. In: Saeed K, Chaki R, Cortesi A, et al. (eds) *Computer information systems and industrial management. CISIM 2013. Lecture notes in computer science, vol 8104*. Berlin, Heidelberg: Springer, 2013, pp. 379–388.
81. Reed R, Oh S and Marks RJ. Regularization using jittered training data. *Proceedings of the 1992 IJCNN (International Joint Conference on Neural Networks)* 1992; **3**: 147–152.
82. Frölich M. Finite-sample properties of propensity-score matching and weighting estimators. *Rev Econ Stat* 2004; **86**: 77–90.
83. Huber M, Lechner M and Wunsch C. The performance of estimators based on the propensity score. *J Econom* 2013; **175**: 1–21.
84. Health and Retirement Study, public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, 2022, https://hrs.isr.umich.edu/about (accessed 1 May 2022).
85. SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015; **373**: 2103–2116.
86. Xu H, Garcia-Ptacek S, Trevisan M, et al. Kidney function, kidney function decline, and the risk of dementia in older adults: a registry-based study. *Neurology* 2021; **96**: E2956–E2965.
87. Shapley LS. A value for n-person games. In: *Contributions to the theory of games, volume II*. Princeton, NJ: Princeton University Press, 1953, pp.307–318.