Received: 2 July 2022 | Revised: 29 March 2023 | Accepted: 24 July 2023

DOI: 10.1111/obr.13635

REVIEW

Weight Management / Interventions

OBESITY Reviews WILEY

Design, analysis, and interpretation of treatment response heterogeneity in personalized nutrition and obesity treatment research

Roger S. Zoh[1] | Bridget H. Esteves[2] | Xiaoxin Yu[1] | Amanda J. Fairchild[3] | Ana I. Vazquez[4] | Andrew G. Chapple[5] | Andrew W. Brown[6] | Brandon George[7] | Derek Gordon[8] | Douglas Landsittel[1] | Gary L. Gadbury[9] | Greg Pavela[10] | Gustavo de los Campos[11] | Luis M. Mestre[12] | David B. Allison[1]

Correspondence
David B. Allison, Indiana University School of Public Health-Bloomington, SPH 116, 1025 E. Seventh Street, Bloomington, IN 47405, USA.
Email: allison@iu.edu

Funding information
This work was supported by the National Institutes of Health under award numbers R25DK099080, P30AG050886, U24AG056053, and R25HL124208 (DBA and AWB); the National Institute on Aging (NIA) under award number 3R01AG057703-02S1 (LMM); and the National Cancer Institute under award number U01-CA057030-29S1 (RZ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Summary

It is increasingly assumed that there is no one-size-fits-all approach to dietary recommendations for the management and treatment of chronic diseases such as obesity. This phenomenon that not all individuals respond uniformly to a given treatment has become an area of research interest given the rise of personalized and precision medicine. To conduct, interpret, and disseminate this research rigorously and with scientific accuracy, however, requires an understanding of treatment response heterogeneity. Here, we define treatment response heterogeneity as it relates to clinical trials, provide statistical guidance for measuring treatment response heterogeneity, and highlight study designs that can quantify treatment response heterogeneity in nutrition and obesity research. Our goal is to educate nutrition and obesity researchers in how to correctly identify and consider treatment response heterogeneity when analyzing data and interpreting results, leading to rigorous and accurate advancements in the field of personalized medicine.

KEYWORDS

heterogeneity of treatment effect, personalized medicine, personalized nutrition, tailored treatment

1 | INTRODUCTION

In clinical trials, one may expect to observe consistent outcomes across an entire treatment group. Even in tightly controlled experimental settings, however, numerous measurable and unmeasurable factors could contribute to differences in outcomes observed among experimental units. Treatment response heterogeneity (TRH), or differences in treatment effects, not merely differences in outcomes among persons, has been explored in nutrition and obesity research with the rise in interest in precision or personalized medicine by clinicians and scientists alike.[1] A

Abbreviations: ADOPT, Accumulating Data to Predict Obesity Treatment; FD, factorial design; MMR, moderated multiple regression; PGRN, Pharmacogenomics Research Network; PRPT, partially randomized preference trial; R2R, randomization-to-randomization; RCT, randomized controlled trial; RRD, repeated randomization design; TRH, treatment response heterogeneity.

For affiliations refer to page 12

goal of characterizing and understanding TRH is to develop more personalized and tailored treatments based on individual characteristics that have a higher likelihood of success.[2] Classically, researchers have relied on secondary analyses of clinical trials to estimate treatment effects for subgroups of individuals that are determined by certain values of genetic, behavioral, or metabolic variables. A treatment-by-subgroup (e.g., sex, genotype, and presence/absence of metabolic disease) interaction can be assessed to determine how treatment effects differ across subgroups. However, individuals can belong to many subgroups, and the reasons for a treatment effect following a dietary change or caloric restriction (or any number of other interventions) cannot be simplified to a single grouping factor. Ignoring these considerations can lead to mischaracterization of treatment effects and will confuse, rather than advance, the burgeoning field of personalized medicine. As the consumer, clinical, and scientific interest in “responder analyses” and precision treatments continues to grow, understanding and implementing proper design and analysis plans to investigate TRH are needed.[3] In this review, we aim to dispel a common misunderstanding of variability in outcome as representing variability in treatment response, describe study designs beyond a parallel randomized controlled trial (RCT) that may offer a better opportunity to disentangle TRH, and outline considerations for statistical analysis. Although we draw on several examples from outside the nutrition and obesity field, similar principles and methodology can be applied to obesity-related research.

2 | DEFINING TRH

2.1 | Change is not response

When assessing the degree of TRH in a study, a precise definition of a *true effect* is needed to distinguish this from a change in an outcome variable from one time point to the next, such as in a pre/post design when two conditions are being compared in a study: a treatment (*T*) and a control (*C*). Suppose that at a particular time, an outcome will be measured on an individual, for example, weight change, which we will denote as *Y*. At that particular time, two possible outcomes for that individual may be considered: the outcome if on treatment *T*, *Y(T)*, and the outcome if on control *C*, *Y(C)*. However, only one of these outcomes will be realized in the data depending on which treatment, *T* or *C*, was assigned to that individual. The other unobserved outcome has been called a counterfactual.[4] The pair of outcomes, *Y(T)* and *Y(C)*, are potential outcomes, and this potential outcomes framework has been referred to as Rubin's causal model after the work of Rubin.[5] The effect of the treatment with respect to the control at that particular time, denoted *D*, is given by *D* = *Y(T)* – *Y(C)*, which we will call the treatment response. A population mean treatment effect is typically estimated in experimental studies. However, our focus herein is TRH, which is how much the variable *D* varies across individuals. However, in the traditional single randomization trial, *D* cannot be observed for any individual, each individual having only been exposed to *T* or *C*. Despite this, much can be learned about the effect of a test treatment versus a control treatment across a group of individuals using controlled trial designs.

One such design is the pre/post study, and we now distinguish *change in outcome* versus *treatment response*, the latter being described above as *D*. The need for such clarification was highlighted by Senn.[6] For example, suppose that at the beginning of a study (i.e., baseline), an outcome such as current weight, denoted as $G(\tau_1)$, is measured in an individual. This individual is then given treatment T for a period of time and the outcome measured again, resulting in a measurement $G(\tau_j)$, the weight of that individual at time *j*. The change in weight from time 1 to time *j* is often interpreted as the response due to treatment.[7] However, the difference $G(\tau_j) - G(\tau_1)$ is what we denoted $Y(T)$ in the above paragraph. The counterfactual outcome, *Y(C)*, is missing. If it is assumed that *change in outcome* is the same as *treatment response*, then the assumption is also made that, had that individual been assigned to the control treatment *C*, the weight change for that individual over that time period would have been zero. But as we cannot know what the weight change of that individual would have been in the control treatment, *C*, this is an unjustified assumption. This assumption creates confusion not only in assessing the degree of TRH but also in simply estimating a mean treatment effect. Information regarding the counterfactual *Y(C)* is needed. Including a control group in a randomized parallel group design can help to estimate this.

2.2 | Control groups or conditions are needed

A true counterfactual cannot be determined, as the same individual cannot simultaneously receive both a treatment and a control condition. The use of an appropriate control group can improve causal inference. At baseline, a group of individuals can be randomized to receive either treatment (*T*) or control (*C*). After a time period, the outcome of interest is measured. Depending on the condition to which an individual is randomized, either *Y(T)* or *Y(C)* will be observable. Although the effect in an individual is still not observable (because of an unobserved *D*), an unbiased estimate of a mean treatment effect can be computed by taking the difference in the averages of observed *Y(T)* and *Y(C)*. We denote this sample average difference by $\overline{d} = \overline{y}(T) - \overline{y}(C)$, where $\overline{y}(T)$ and $\overline{y}(C)$) are the sample means of the outcome variables for the treatment and control groups, respectively. The quantity $\overline{d}$ is an unbiased estimate of a population mean effect, denoted here as μ_D for the parameter and $\widehat{\mu}_D$ for the sample estimator. No assumption was needed (nor would be appropriate) to equate a change over time with treatment response.

3 | TESTS FOR TRH

There remains a challenge in assessing the degree of TRH in a study. We denote a measure of this heterogeneity by the variance of individual treatment effects, or the variance of *D*, represented by σ_D^2 for a population and s_D^2 in a sample of individuals. The sample quantity, s_D^2, cannot be computed from observed data because these data contain no information to compute the sample correlation between the two values *Y(T)* and *Y(C)*. Because the two potential outcomes are not observable on any individual, there is no way to compute an estimate of σ_D^2. This issue and ways to address it have been discussed elsewhere.[8,9] Herein, we focus on some key results shown in these references and how they relate to a two-group parallel randomized design.

Although s_D^2 cannot be computed from observed data, the sample variance of outcomes in the two groups can be computed and are denoted as $s_{y(T)}^2$ and $s_{y(C)}^2$. The corresponding population quantities would be denoted by σ replacing s. Williamson et al.[7] argue in favor of including the observed values in the control group in assessing whether there is significant variability across individual values due to treatment. This can be done by using the difference $s_{y(T)}^2 - s_{y(C)}^2$ (or ratio $s_{y(T)}^2/s_{y(C)}^2$) to determine whether there is more (or less) variability in observed individual outcomes in the treatment group than in the control group. If $s_{y(T)}^2$ is greater than $s_{y(C)}^2$, and this difference (ratio) is significantly different from zero (from one), then there is evidence that individual outcomes vary more under *T* versus *C* in a broader hypothetical population from which individuals were sampled. Potential causes of this larger variation could be investigated, such as individual attributes or covariates that may explain why, in the case of weight change, some individuals lost more weight than others within the treatment group. Williamson et al.[7] are mostly careful to refer to *Y(T)* and *Y(C)* as "changes" rather than "responses." However, the improper use of *response* when referring to a change in outcome is still prevalent in nutrition and obesity literature. Such incorrect terminology can add to confusion regarding the effectiveness of a treatment across individuals.[6]

3.1 | TRH variance can be bounded

As a correlation is constrained to the interval −1 to 1, a lower and upper bound for the variance of *D* (s_D^2) can be computed, and thus, a lower and upper bound for σ_D^2 can be estimated.[8,10] The lower bound for the sample variance s_D^2 is given by $\left(s_{y(T)} - s_{y(C)}\right)^2$, which occurs when the sample correlation between the two potential outcomes is equal to 1. An upper bound for s_D^2 is given by $\left(s_{y(T)} + s_{y(C)}\right)^2$, which occurs when the sample correlation is equal to −1. If the difference $s_{y(T)} - s_{y(C)}$ are statistically different from zero, then it is a reasonable conclusion that some TRH exists in the population.[7] The degree of this heterogeneity could be even greater than that indicated by the quantity $\left(s_{y(T)} - s_{y(C)}\right)^2$, but it would not exceed that given by the quantity $\left(s_{y(T)} + s_{y(C)}\right)^2$ (barring random sampling variations). In clinical studies, it may be more reasonable to assume that the correlation between potential outcomes *Y(T)* and *Y(C)* is closer to 1 than −1, but there is not a way to test this assumption from typically observed data. One approach to testing whether $\sigma_D^2 > 0$ is the *F* test for equal population variances. An interesting conclusion, then, from the results discussed here is that the same *F* test is also a test for the presence of TRH. If the result of such an *F* test is significant, then it may be worthwhile to investigate further whether other variables, that is, covariates, may explain why some individuals respond differently to treatment *T* versus *C*. Furthermore, if there is evidence that the degree of treatment heterogeneity in a population is large, then the mean treatment effect may be misleading when evaluating the effect of treatment *T* versus *C* in a population.[7,9] We note that presence of TRH does not always manifest itself in terms of an increase variance in the treatment group compared to the control group.[11,12] In fact, Senn[13] (in Section 3.2) shows a striking example where heteroscedasticity is present between the positive control group and treatment group. Yet, the variance is reduced as the treatment effect increases. Thus, the *F* test for equal population variances is appropriately a two-sided *F* test that reject the null hypothesis of equal variances for small or large variance ratios.

Thus far, we have considered *Y(T)* and *Y(C)* to be continuous outcomes (e.g., weight change). Different techniques are needed when evaluating TRH with binary outcomes (e.g., success vs. failure).[14,15] Regardless of whether the outcomes are binary or continuous, the challenge in trying to accurately evaluate TRH results from *Y(T)* and *Y(C)* not being observable together for an individual in a two-group parallel study. In fact, at any particular time point, *Y(T)* and *Y(C)* can never be observed together for an individual, what Paul Holland named the "fundamental problem of causal inference."[16] However, if circumstances allow, a different randomized trial design may be employed that, when combined with some less onerous assumptions that may be reasonable depending on the treatment under study, a "value" for *Y(T)* and *Y(C)* may be "available" for an individual, thus allowing observation of a quantity that represents *D*, the individual effect of treatment *T* versus *C*. Examples discussed in this review include crossover, Balaam, continuous-dose designs, and Loop designs.

3.2 | Assessing TRH as a function of observed scalar prerandomization covariates: Good old-fashioned moderator analysis

Despite the plethora of advanced options available to test for TRH, there are several scenarios where a simpler approach, "good old-fashioned moderator analysis," may suffice. Limiting the number of moderator variables examined in these scenarios is paramount so that the analyses remain manageable, because the inclusion of numerous moderators in this approach can become unwieldy quickly with the need to interpret higher-order interaction terms and diminished power to detect individual effects.

Examining TRH as a function of observed scalar prerandomization covariates may be particularly useful in iterative approaches to improving interventions for vulnerable and underserved populations, where moderation and mediation analyses are considered in tandem.[17] By identifying putative moderators, such as race/ethnicity or socioeconomic status, a priori and outlining interactive effects in subsequent analysis of primary outcomes, it becomes possible to empirically evaluate underlying mechanisms of key disparities observed. This approach has the potential to mitigate disparities by both lending insight into any needed implementation refinements for these groups, as well as by potentially identifying alternative behavioral determinants specific to the underserved populations.

Another example of when moderator analysis may be a judicious strategy to examine TRH occurs when there is theory to suggest effect modifiers at the outset of a study. Baseline-by-treatment interactions represent one such scenario, in which treatment effects depend on an individual's initial status.[18] Interactive effects of this kind may manifest in baseline severity leading to either greater or lesser treatment responsiveness.[19] For example, a baseline-by-treatment interaction was demonstrated in an RCT in patients with type 2 diabetes.[20] Study results showed that participants at higher risk benefited more from a combination drug therapy targeting glycemic control than did individuals with lower HbA1c levels at baseline.

One approach to determining TRH from moderator analysis is moderated multiple regression (MMR) analysis. MMR investigates heterogeneous treatment effects via the general, or generalized, linear model by incorporating an interaction term between the treatment indicator and a pretreatment covariate. Presuming random assignment, assumptions include that putative moderators are measured before treatment and do not correlate with treatment allocation.[21] The framework permits evaluation of both categorical and continuous covariates, with the former requiring an appropriate coding scheme of the variable, such as dummy or effect coding.[22] The simplest MMR model is given by:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 Z_i + \beta_3 T_i Z_i + \varepsilon_i,$$

where Y_i is the outcome for individual *i*, β_0 is the model intercept, β_1 is the partial regression coefficient corresponding to the effect of treatment (*T*), β_2 is the partial regression coefficient corresponding to the effect of the pretreatment covariate (*Z*), β_3 is the partial regression coefficient corresponding to the interactive effect (*TZ*) of treatment and the pretreatment covariate, and ε_i is the residual in the model distributed as independent and identically distributed (*i.i.d.*) with $N(0,\sigma^2)$. Modeling lower-order terms ensures that effects are not confounded and that the test of significance for the interaction is accurate, as estimates of variance for model parameters are not orthogonal. Mean centering variables in the model can facilitate interpretation of lower-order terms.[23]

A test of statistical significance for TRH is given by the *t* statistic for β_3. Statistical significance of β_3 indicates the model is nonadditive, such that treatment effects are conditional on levels of the pretreatment covariate. Researchers can alternatively use nested models (i.e., $R^2\ \Delta$) to test for heterogeneous treatment in MMR:

$$F = \frac{\left(R_2^2 - R_1^2\right)/(k_2 - k_1)}{\left(1 - R_2^2\right)/(N - k_2 - 1)},$$

where R_2^2 is the R^2 associated with a model that includes both main effects and the interaction and R_1^2 is the R^2 associated with a main-effects-only model. The k_2 are the degrees of freedom of the model that includes both main effects and the interaction, and k_1 are the degrees of freedom of the model that includes only main effects. The square root of the *F* statistic from the $R^2\ \Delta$ test will equal the absolute value of the *t*-test statistics for β_3. For tests involving one degree of freedom, the results of the *F* test are equivalent to that of a *t* test; however, we focus on the *F* test because it is more general (i.e., when $k_2 - k_1 > 1$).

Graphing simple slopes of the treatment effect at different values of the covariate lends insight into how the slope changes in the presence of a significant interaction:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_2 Z + \left(\widehat{\beta}_1 + \widehat{\beta}_3 Z\right) T.$$

When the pretreatment covariate is categorical, the user may plot simple effects of treatment at values of *Z* that code group membership. When the pretreatment covariate is continuous, it is useful to plot simple effects at $-1sd$ below the mean of Z, $+1sd$ above the mean of *Z*, and at the mean of *Z* to characterize effect heterogeneity across a wide range of data in which most observed scores will fall (assuming normally distributed data).

Finally, computing an effect size for the interaction term in the MMR model will help to portray the practical significance of the effect. Variance-explained measures are most used, and f^2 gives the variance explained by the interaction above and beyond lower-order effects in the model:

$$f^2 = \frac{r_{Y.MI}^2 - r_{Y.M}^2}{1 - r_{Y.MI}^2},$$

where $r_{Y.MI}^2$ is the squared multiple correlation associated with the full MMR model and $r_{Y.M}^2$ is the squared multiple correlation associated with a main effects model that does not include the interaction.

In the case of continuous outcome, estimation and inference of multiplicative interaction are easily implemented in any software. However, in the case of binary outcome, in addition to the multiplicative interaction, the additive interaction can also be considered.[24–27] It is important to mention that the presence or absence of interaction can sometimes be scale dependent (multiplicative vs. additive), which is also related to the idea of ordinal versus disordinal or rank versus non-rank interaction.[28,29]

4 | STUDY DESIGNS

Previous sections of this review have outlined why conventional trial designs, such as the two-group parallel RCT, cannot estimate the total amount of TRH. We have outlined that participants in research interventions can be partitioned into an infinite number of "subgroups" based on both measurable and nonmeasurable covariates. Conducting analyses based on one or even multiple of these covariates cannot give an accurate estimation of all sources of TRH.[30] Additionally, a change (in the outcome of interest) is not the same as a response to an intervention. Therefore, this section will outline trial designs in which the participant is exposed to both control and intervention treatments.

4.1 | Factorial designs (FDs): Assessing whether treatments affect other treatments' effects

It is possible that the response of an individual to a treatment may depend on previous exposures, including whether the individual has been exposed to another treatment. A potential mitigation of these effects is the FD. FD measures the effects of more than one experimental factor in the same study. An example is a study that includes two factors, each with two levels, called a 2-by-2 factorial. For example, an experiment included the presence or absence of sucralose as one factor and the presence or absence of sucrose as another factor,[31] resulting in four treatment groups: neither sucralose nor sucrose; sucralose with no sucrose; no sucralose but with sucrose; and both sucralose and sucrose. Other studies have included exercise as one factor and diet as another or group or individual treatment as one factor and personal preference for treatment modality as another.[32]

FD permits the experimenter to determine whether the effects of one factor (e.g., sucralose) are different in the presence of another (e.g., sucrose). When such a difference exists, the factors are said to interact, which is determined through a formal test for an interaction effect, commonly through a two-way analysis of variance (ANOVA). An interaction means that the outcome in the presence of both factors A and B together is not just the addition of the effects of factors A and B alone. In FD, it is essential to test for interaction effects before interpreting the effects of an individual factor. When an interaction effect exists, the interpretation of the effects of one factor is dependent on the level of the other factor, and so averaging across the two factors (as done in estimating the "main effects" of one factor at a time) may be inappropriate. Instead, results are often discussed in terms of "simple effects" when there is an interaction effect, which are within-factor comparisons (e.g., with and without sucralose) while holding levels of the other factor constant (e.g., when sucrose is or is not provided). These effects are depicted in Figure 1.

The lack of an interaction effect implies that the level of the other factor does not matter, and the two factors can be treated as independent effects. In the 2-by-2 example described above, this means there are two main effects: the difference in outcome in the presence or absence of sucrose, each averaged over levels of sucralose, and the difference in outcome in the presence or absence of sucralose, each averaged over levels of sucrose. The lack of an interaction effect implies that the effects of the levels of one factor do not depend on the levels of other factors. The advantage of FD when there is no interaction effect is efficiency: When the design is balanced, the study can test two independent factors at the same time with twice as many participants for each factor.

The 2-by-2 FD concept can be generalized to situations in which one of the factors is not experimentally assigned. Consider a study in which the effects of a low-carbohydrate diet are thought to be different from those of a low-fat diet (factor A) depending on whether the participant has a particular genotype thought to be associated with carbohydrate metabolism.[33] Although diet was assigned, genotype was not. This applies to testing for differences between race and sex as well. FD can be generalized to experimental factors that are categorical (described above, e.g., sucrose vs. sucralose), discrete (e.g., number of treatment sessions in a treatment), or continuous (e.g., percent of energy consumed as carbohydrate) and can be used in a variety of research, including enzyme kinetics,[34] in vitro chemical research,[35] and animal studies.[36]

There are statistical factors to consider when selecting FD. In many (but not all) situations, the power for an interaction test (INT) is much lower than the power for main effects.[37,38] When the

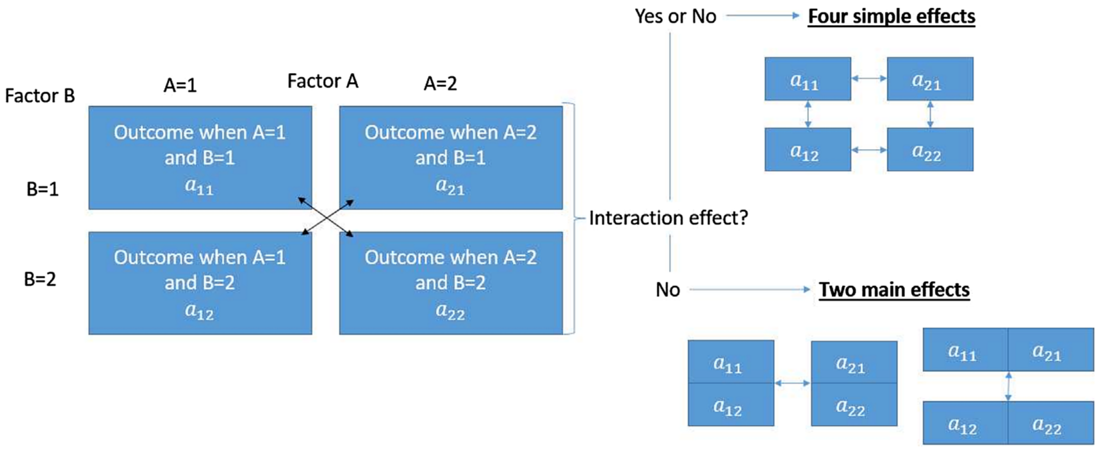
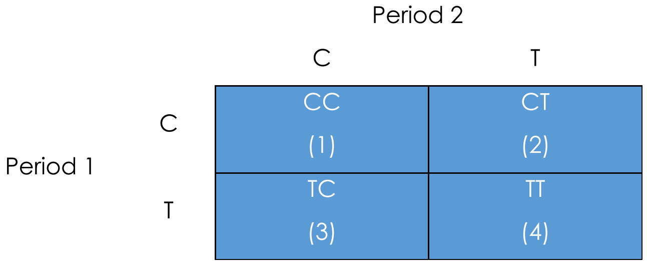


FIGURE 1 Examples of interaction effects, main effects, and simple effects for a 2 × 2 factorial design. This example assumes two factors (A and B) each with two levels (1 and 2). Consider from the text the examples of factor A representing sucralose, with level 1 being no sucralose and level 2 being sucralose, and similarly for factor B representing sucrose, with level 1 being no sucrose and level 2 being sucrose. Interaction effects should be formally tested before attempting to test for or interpret main effects because by definition the effect of one factor depends on the effect of another. However, simple effects can be tested whether or not there is an interaction effect. If the researcher is willing to accept there is no interaction effect, then main effects are more highly powered. See Figure 2 for examples of interaction effects.



		Period 2	
		C	T
Period 1	C	CC (1)	CT (2)
	T	TC (3)	TT (4)

FIGURE 2 The Balaam design with treatment sequence CC, CT, TC, and TT. Individuals are randomized to each of the 4 treatment sequences.

interaction term is of major interest, the study should be powered to test the interaction test. Because the main effects are averaged over the levels of the other factor, the main effects have twice as many observations in a balanced design than do any simple effects or interaction effects. Next, as described above, interpreting main effects without considering whether there is an interaction effect risks misestimating the effect size of the treatments, which depend on the level of the other factor. If a study is designed as a factorial, then removing an interaction effect from the model should be done with care. Removing it from the model effectively accepts the null hypothesis and therefore implicitly changes the model from testing whether an interaction exists to declaring that the two factors can be treated as operating independently. Selection of model terms that are based on theory may be appropriate, that is, maintaining the 2-way ANOVA and including interaction effects, similar to how *p*-value-based selection of variables through stepwise regression has fallen out of favor.[39]

An FD is relevant to TRH when the heterogeneity of treatment effects can be determined by some other unmeasured patient factor (i.e., dietary or exercise regimens). Here, it may appear that TRH exists, when really the TRH is driven by a separate factor shared by some patients. However, if these factors are actually measured, it may be possible to find that TRH exists within levels of that measured factor. If individuals are treated with a pharmaceutical, what is the incremental effect of adding an intensive lifestyle intervention? When one of the factors is an inherent characteristic (e.g., race and sex), FD can test whether the effects are different depending on that prognostic factor using a generalization of FD. Such cases are limited in causation because the factor is observed rather than assigned (e.g., whether race is the cause of the difference or whether factors associated with race caused the difference). Nonetheless, FD can identify factors in which the expected effects might differ among levels of a factor. We further note that even within a particular assigned level of each of the two treatments in the 2-by-2 FD, there could still be individual TRH that cannot be directly estimated.

4.2 | Crossover designs

This design randomizes a sequence of treatments to an individual, such that the individual will receive all treatments (*T* and *C*) over a sequence of several cycles with the goal of studying differences between treatments.[40] By receiving both *T* and *C*, individual treatment effects can be estimated (technically, this may depend on additional assumptions; see Senn[40] for more details), while also estimating an average treatment effect. As discussed above, unlike trials in which an individual receives only one condition, crossover designs can estimate individual effects directly because of multiple measurements on each patient in both *T* and *C* conditions. The variance of the estimated individual treatment effects can therefore be used to quantify the degree of TRH. Despite this benefit of crossover designs, very few have conducted assessments of treatment heterogeneity when analyzing data from crossover trials.

Gewandter et al.[41] assessed TRH in four crossover designs for fentanyl's effectiveness to reduce episodes of extreme pain in cancer patients. Following the approach suggested by Ezzet and Whitehead,[42] the researchers used a mixed-effects model with fixed effects for treatment and random effects for each patient and treatment–patient interaction for binary data. By considering both a patient-level random effect and a treatment–patient interaction, patient heterogeneity on outcome can be separated from the treatment heterogeneity seen in patients. Random effect models[43,44] also provide benefit in that the variance of these effects is estimated and can be used in hypothesis testing. These researchers additionally investigated whether the patient-level random effect variance was 0 (i.e., no treatment heterogeneity) and found that three of the four studies had significant treatment heterogeneity for fentanyl.[41] The authors suggested that this approach could be used in other crossover designs, which might warrant future genomic or biomarker studies to determine whether these factors are driving the apparent TRH, thus leading toward more personalized medicine.

Other models for estimating TRH in crossover designs that have yet to become popular are Bayesian hierarchical models,[45,46] which assume that each patient has their own treatment effect, denoted μ_i, that is drawn from a (Gaussian) distribution centered around some grand treatment effect μ_0. Several authors have used Bayesian hierarchical models to estimate a shared treatment effect from pooled n-of-1 trials.[47–50] By pooling these n-of-1 trials together, with varying numbers of cycles and treatment sequences for each patient, the authors were able to create effects of a crossover design. Another example of Bayesian hierarchical models in n-of-1 trials comes from Mitchell et al.,[48] who sought to determine whether methylphenidate (MPH) provided a useful treatment of fatigue in 43 Australian adults with advanced stage cancer with at most three cycles of treatment and placebo. The authors assumed that the difference in fatigue scores for each day, denoted Y_{ic}, had a normal probability distribution, with mean μ_i and variance σ_i^2. μ_i is therefore the individual treatment effect of patient *i*, which is assumed to have a normal distribution with mean μ_0, which represents the overall treatment effect.

Another study design that can allow for estimation of individual treatment effects is the Balaam design. The Balaam design is considered when there is evidence to suggest that the treatment response of the experimental unit (for example, an individual in a clinical dietary trial) may vary depending on the sequence in which treatments are assigned. For example, in a crossover design, the true effect of treatment *T* may be impacted by the individual having been exposed to *C* first in addition to or in the absence of a treatment *C* residual effect.

The Balaam design, first introduced as a two-period design with *t* treatment levels and t^2 experimental units, refers to a study design in which each of the experimental units in the study is randomly assigned to a pair or sequence of treatments.[51] In this design, individuals are randomized each of the t^2 treatment sequences. Unlike the traditional crossover design in which experimental units will absolutely receive both treatments, the Balaam design allows for cases in which sample units receive the same treatment at both periods (CC or TT). Balaam outlines a very simple randomization procedure of the experimental units in the treatment sequence.[51] First, the *t* treatments are randomly assigned unique labels and then the experimental units are randomly assigned to one of the t^2 treatment sequences.

From a statistical viewpoint, the Balaam design has appealing characteristics because it yields treatment effect estimates unbiased by the presence of period-by-treatment effects.[40,52] Statistical optimality (in terms of lower variance of the treatment effect estimates when compared with a parallel design) is attained when the design is balanced, and an equal number of individuals is assigned to each of the treatment sequences. The period-by-treatment effect and the treatment residual effects cannot, unfortunately, be disentangled in this case, and thus, there are limitations to this design. Carryover effect is at least a manifestation of a TRH (i.e., that the effect of treatment differs as a result of preceding treatment), and an accurate estimation of such effect is vital to an appropriate analysis.[52,53] Unfortunately, treatment effect estimates in the Balaam design can be suboptimal (larger variance) compared with the traditional crossover design even in the absence of carryover effect.[53] Mori and Kano[54] proposed a new treatment effect estimate, and an associated test that they demonstrate is very powerful in detecting the presence of a treatment effect in a Balaam design. Another limitation is the often-restrictive assumption of constant carryover effect routinely assumed in the Balaam design. Candel[53] extended the Balaam design to accommodate and provide means to estimate and test a broad form of assumed carryover mechanisms. This has a direct application in pharmacological studies, where pharmacological carryover between placebo and treatment or treatment to another treatment can be complex.[52,53]

Finally, in a Ballam design, we believe that under (moderate) certain assumptions, we can estimate the variance of TRH. This will be followed elsewhere.

4.3 | The Loop design and other $n = 1$ trials

In the Balaam design discussed above, an individual is randomized to receive one of four allocations: *TT*, *TC*, *CC*, or *CT*.[51] However, like the two-period crossover design (where an individual would receive either *TC* or *CT*), this design is limited in its ability to separate the true role of the intervention from other sources of variability in determining individual response. To solve this, Loop et al. proposed the repeated randomization design (RRD), which is similar to the Balaam design, but with a greater number of randomizations and study periods.[55] With multiple randomization periods, multiple observations on the individual's response to both the treatment and control conditions are gathered. This allows for estimation of the mean treatment effect for the individual and the variance of this mean treatment effect.[56] Additionally, by studying individual covariates, we can also understand the characteristics associated with a more desirable response to treatment. This design therefore can provide a more accurate estimate of TRH than traditional RCTs.

The RRD trial can be described as a type of $n = 1$ study, in which subjects are studied on an individual level. Multiple observations are made on the individual, in either an observational setting or an interventional setting. In an observational $n = 1$ trial, data are gathered on the individual over time, without introduction of an intervention. When designing this kind of trial, consideration is needed of how participants' behaviors may change during repeated monitoring.[57] A run-in observation period may be necessary to ensure participants' familiarity with data collection. For interventional $n = 1$ trials, participants may be subjected to treatment or control conditions alternately or through RRD.

These $n = 1$ trials allow for investigation into the effect of various treatments on an individual, identification of the more efficacious treatment for an individual, and examination of the stability of response over time. When considering interventional $n = 1$ trials in nutrition and obesity research, it is crucial to consider the intervention being investigated. Some treatments may not be suited for this type of trial, such as those with long carryover effects or that are not reversible (certain pharmaceuticals, surgery, etc.). However, interventions like use of dietary supplements or single meal challenges may be appropriate. A washout period, during which patient treatment is stopped, between intervention periods may be necessary to mitigate any carryover effects.

4.4 | Assessing whether choosing treatment affects effects: Choice and preference designs

Random allocation of individuals to treatment groups ensures that group differences in all known and unknown confounders are due to chance and that statistical inferences made regarding treatment effects have a valid false-positive error rate.[58] This fact is central to the value of randomized trials but is at odds with how treatments are used in the real world. Treatments are not typically selected via random assignment outside of clinical trials. Many treatments are the result of a deliberate choice by a free-living individual. For example, the decisions to increase consumption of yogurt for its purported probiotic benefits, to skip (or not skip) breakfast as part of a weight-loss diet, or to initiate an exercise program are all made with a high degree of certainty that one is actually consuming yogurt, skipping breakfast, or exercising. Implicit in each of these decisions is a preference for that treatment over another treatment (or nontreatment). Individuals interested in the benefits of probiotics but with a distaste for yogurt would likely prefer a non-yogurt treatment. Conversely, individuals who enjoy yogurt might gladly choose it as a means to increase their consumption of probiotics. Given an individual's infinitude of preferences, randomly assigning a treatment makes it likely that some individuals will be assigned a preferred treatment whereas others are not, potentially leading to "resentful demoralization" or other thoughts and feelings affecting response outcomes.[59]

A mismatch between treatment assignment and preference may introduce heterogeneity in treatment responses in nutrition and obesity research—perhaps with larger treatment effects among individuals randomized to their preferred assignment. Several experimental designs may account for participant preferences and have been used in personalized nutrition and obesity research to examine the effect of treatment preferences. The potential for patient preferences to affect treatment effects was recognized by Brewin and Bradley,[60] who stated, "If effectiveness is evaluated after random administration to patients who may or may not desire the treatment, it will be difficult to distinguish between a treatment that failed because it was not inherently effective and one that failed because it was not targeted towards patients who understood why that treatment was given or who were suitably motivated." This critique is readily extended to randomized trials in personalized nutrition and obesity research: Do weight-loss studies that fail to find a statistically significant treatment effect fail because the weight-loss treatment is generally ineffective, or because it was not sufficiently tailored to the study population or met with sufficient motivation? Participants randomized to a behavioral weight-loss intervention that cannot be blinded, and in which they may not be motivated to participate once assignment is known, may be less adherent and more likely to drop out, complicating interpretations of the treatment effect.[61]

To address such concerns generally, Brewin et al. proposed the partially randomized preference trial (PRPT),[60] also called the parallel randomized and nonrandomized trial[62] or comprehensive cohort design.[63] In this design, participant preferences are assessed before randomization. Those without a preference are randomized to treatment or control, whereas those with a preference are allowed to select their preferred group. Motivational factors and other selection effects could then be assessed by comparing the strength of the treatment effect observed in those randomly assigned to the treatment with the association observed in participants who selected the treatment. In the *fully randomized preference trial*, proposed and implemented by Torgerson et al.,[64] in a trial of exercise for back pain, participant preferences are assessed before randomization, and *all* consenting participants are randomized regardless of preferences, allowing researchers to consider patient preferences in statistical analyses.

The PRPT design is related to several other "hybrid" designs, so termed because they accommodate patient preferences by incorporating both randomized and nonrandomized groups.[62] The *Zelen randomized consent design*, originally proposed to facilitate the recruitment of participants into randomized trials, is distinctive because it randomizes participants *prior* to consent.[65] In the single-consent version of the Zelen design, participants randomized to the experimental treatment are informed of their assignment and given the option to decline treatment. Those randomized to the control group are not so informed. In the double-consent Zelen design, individuals randomized to the treatment group *as well as* individuals randomized to control groups are informed of their assignment and given the option to decline their assigned group in favor of their preferred group.[66] Finally, in the *preference option randomized design*, participants complete the informed consent process prior to randomization (contra the Zelen design and consistent with widely accepted practice); however, during the informed consent process, participants are told that, following randomization, they will be given the opportunity to join a group other than the one they were randomized to if they "clearly express their preference" for doing so. Otherwise, they will remain in the group to which they were randomly assigned.[67]

Given ethical concerns with the Zelen design, and the additional complexity associated with other hybrid designs (e.g., Rucker[68]), it is valid to ask whether evidence exists that patient preferences affect treatment effects both generally and in nutrition and obesity research specifically. A 2005 systematic review and meta-analysis of RCTs incorporating participant preferences identified 27 comprehensive cohort designs and 5 doubly randomized designs published between 1966 and 2004.[63] This review concluded that recruitment was affected by participant preferences and that more highly educated individuals were more likely to refuse randomization. However, there was little evidence that patient preferences had large effects on outcomes and no evidence that preferences affected attrition, concluding that this review "… gives less support to the hypothesis that preferences significantly compromise internal validity."[63] Another systematic review and meta-analysis of eight fully randomized patient preference trials similarly concluded that participant preferences did not have an effect on attrition but found larger effects of preferences on outcomes, such that estimates of the treatment effect were larger among participants randomized to their preferred treatment compared with participants indifferent to group assignment or those assigned to a group other than their preferred treatment.[69]

In dietary weight-loss interventions, William et al.[70] developed a doubly randomized preference trial protocol to assess the role of dietary preferences in weight-loss trials and later reported that allowing individuals to choose their diet did not cause greater weight loss or dietary adherence.[71] In an earlier study of participant dietary preferences, the PREFER study, researchers found that mean percentage weight loss was greater among individuals randomly assigned to a diet than in those assigned to their preferred diet.[72] A 2019 systematic review and meta-analysis of nine studies investigating the role of participant choice in weight-loss strategies, including the two discussed above, found no evidence that individuals assigned to a preference diet lost more weight or dropped out in higher numbers than those randomly assigned a diet. Thus, while allowing participants to choose their treatment might intuitively be thought to affect effects; to date, there is little evidence that substantial treatment heterogeneity is introduced by mismatched preferences in dietary weight-loss trials.

4.5 | Assessing whether expectancies affect effects

As discussed above, although traditional RCTs are the mainstay for estimating the causal effect of treatments, they usually do not reflect how administration of the treatment would occur "under the conditions of its intended use."[73] This is because in actual use, individuals receiving a treatment, taking a drug, eating yogurt, skipping breakfast, exercising, and so on are not blinded and are thus subject to expectancy effects. An individual's knowledge of enrollment in the treatment or control condition can lead to treatment-related outcome expectations and may impact the treatment in several ways. This can be due to conscious action; for example, an individual enrolled in a weight-loss study may make additional lifestyle modifications to lose weight if he or she strongly feels that the treatment will be beneficial. These expectancies also may happen more subtly, such as in a study by Elliman et al.[74] that found that providing subjects with caffeine improved performance on a vigilance test only if the subjects were told that they had been given caffeine. In either case, expectancies should be of interest to researchers because when these expectancies interact with the treatment, they can produce biased estimates of the treatment effect under the actual conditions of use when those estimates are from conventional blinded RCTs.

One method for estimating expectancy-by-treatment interaction effects involves the use of the "balanced-placebo design,"[75] in which a 2 ×2 design matrix is created wherein subjects are randomized on two factors: to receive treatment or control and to be told that they were given treatment or control, independent of what they were actually given. Although this study design can distinguish expectancy effects from treatment effects and can identify any possible interactions, it carries ethical concerns because it involves participant deception. This may be acceptable in short-term studies of healthy subjects who have consented to deception, but there would be ethical concerns for such deception of patients seeking treatment for an illness.[76] Additionally, it may not be feasible for certain nutrition or obesity interventions, in which subjects are given a specific food or enrolled in an intensive intervention that cannot be blinded.

An alternative to the balanced-placebo design that does not involve participant deception is the randomization-to-randomization (R2R) study design.[77] This design works to modify subject expectancies by the following process:

1. Subjects are randomly assigned a probability p of receiving treatment between 0 and 1 (not including 0 or 1).
2. The subject is told their probability.
3. The subject is assigned to receive treatment or placebo based on their given probability p, while maintaining blinding.
4. The trial is conducted as usual, and in the data analysis, the assigned probability and a treatment-by-probability interaction term are included in the appropriate analytic model.

The choice of distribution to generate values for p in Step 1 should be made in such a way that it provides desired coverage over the range of probabilities between 0 and 1. For Step 3, the law of large numbers dictates that one should obtain approximately the proportions of subjects in each treatment group as the probability indicates (e.g., 3 of 10 subjects given a probability of $p = 0.3$ in the treatment arm) with a large sample size. However, one can safeguard against imbalances with small samples by making assigned probabilities discrete and finite in the randomization to both the probability (e.g., 10% of the sample to $p = 0.1$, 10% to $p = 0.2$, 10% to $p = 0.3$, and so on) and the treatment (e.g., of the 10% assigned to $p = 0.3$, 30% receive treatment, and the other 70% placebo).

For the analysis of the R2R study design, expectancies can be incorporated in the following linear model:

$$Y = \beta_0 + \beta_1 T + \beta_2 p + \beta_3 T * p + \varepsilon$$

Using this model, one can calculate not only the unique contributions of treatment and expectancy effects and their interaction but also the treatment effect under "actual conditions of use" as the quantity $\beta_1 + \beta_3$. This represents the difference between an individual receiving the placebo ($T = 0$) and another receiving the treatment ($T = 1$) when both would fully expect to receive treatment ($p = 1$). The balanced-placebo design can also be viewed and analyzed in such a framework, wherein the subjects are assigned to only $p = 0$ and $p = 1$ and that assignment is independent of the actual treatment received. TRH comes in when we consider the interaction effect. A statistically significant $\beta_3 = 0$ implies that the effect of treatment is heterogeneous over levels of subject expectancy.

When considering these two designs, it is important to also consider their limitations. The balanced-placebo design allows for direct estimation of the treatment effect under actual conditions of use, whereas the R2R design requires the analyst to extrapolate to $p = 1$ from the assigned probabilities, which requires specifying the shape of the effects of expectancies (linear or otherwise) and a sufficient sample size near $p = 1$ to make a reasonable estimate. Although the R2R design is more challenging in the execution and analysis than the balanced-placebo design, it does not carry the ethical concerns of subject deception that may disqualify the balanced-placebo design from use in many contexts.

5 | SPECIFIC APPLICATION OF TRH: ASSESSING TRH AS A FUNCTION OF GENETIC PREDICTORS

When selecting potential prerandomization covariates, one can consider both genetic and environmental factors. Genetic factors can affect disease predisposition, disease progression, and response to treatment. There are at least two mechanisms by which genetic factors can affect TRH. First, genes can affect drug or nutrient absorption, metabolism, and excretion, thus affecting the effectiveness of treatment and adverse reactions. For example, several variants within genes coding for enzymes of the cytochrome P450 family have been reported to affect drug metabolization and thus TRH[78,79] and drug–

drug interactions.[80] Hypersensitivity reactions to drugs have also been shown to have a genetic basis and to be associated with genetic variants in the human leukocyte antigen system.[81]

Complex diseases including cancer, type 2 diabetes, and obesity are the result of multiple genetic variations that are difficult to elucidate from symptoms and clinical phenotypes. Genomic data can help to identify these subtypes, such as in groups of patients with variations in a specific pathway that may require targeted treatments (to which patients will respond differently). Therefore, a second way by which genetic information can inform likely TRH is by defining subtypes and their underlying genetic etiology, which can be used to design and prescribe targeted treatments. For example, gene expression data have been used to classify breast cancer tumors into subtypes[82,83] and to apply targeted therapies to each of the subtypes (i.e., personalized medicine).[82,84]

The preceding examples illustrate how genetic information may be informative about TRH. However, identifying the genetic factors affecting response has been (and remains) challenging. As interest in genetic-based health recommendations and personalized medicine grows, it is no surprise that the number of genome-wide association studies published continued to rise. These studies have associated thousands of genetic variants with many different complex traits, including obesity, cancer risk, and type 2 diabetes.[85] However, only a small fraction (less than 10%) of genetic association studies have investigated the role of genetic factors in TRH.[86] Many of these discoveries have originated from candidate gene studies, often informed by genome-wide association studies, that have targeted large-effect genes physiologically involved in drug metabolism and immune response. Regarding other complex mechanisms, such as physiology of weight gain or loss, there is much work to be done.

Low power may be the single most important factor that has limited the ability to identify genetic variants associated with TRH. For complex traits and diseases, individual variants often explain a small fraction (e.g., less than 0.1%) of genetic risk; therefore, a large sample size (e.g., hundreds of thousands) is required to achieve moderate power. When investigating outcomes following a clinical trial, the challenge may be even greater because of the many possible treatments for chronic diseases and wide heterogeneity in patient outcomes. Therefore, a very large initial sample size is required to achieve high power, much larger than most clinical trials, which are often not powered to detect genetic effects. To confront limited power, several consortia have been formed. The Pharmacogenomics Research Network (PGRN, https://www.pgrn.org/) was formed to catalyze and lead research and translation on the genetic basis of TRH and in pharmaceutical trials. There are similar consortia being formed within the field of nutrition and obesity, such as the Accumulating Data to Predict Obesity Treatment (ADOPT) Project[87] and the Trans-NIH Genetics Consortium.[88,89]

Efforts such as the PGRN, ADOPT, and the Trans-NIH Genetics Consortium will help to further advance the study of the genetic basis of TRH; however, with the current sample sizes, power remains limited. This implies that most small-effect variants (which collectively explain a sizable fraction of interindividual differences in TRH) may remain undetected, because the false-positive rate may be high. Complex-trait genetic studies suggest that incorporating variants that individually may not reach genome-wide association significance into polygenic risk scores[90] and whole-genome regressions[91] may improve prediction accuracy. Polygenic prediction methods can help to advance the detection of TRH; however, accurate prediction will require much larger sample sizes than those currently available.

Absence of randomization and sampling bias are other important challenges for genetic studies of TRH. Many of the modern multi-omic data sets are observational studies, either population prospective cohorts or data repositories of samples and data from patients. However, treatment assignment in these studies was often guided by genomic information (e.g., gene expression data may have been used to prescribe therapies to cancer patients). This can cause biases on inferences of differential TRH. Well-designed clinical trials can provide unbiased estimations of TRH; however, these trials can be expensive, lengthy, and require many participants to obtain adequate power. Therefore, an effective strategy may require a combination of the use of data from observational studies and RCTs.

6 | DISCUSSION

We are all special and unique, just like everybody else.* Yet, on the other hand, as Daniel Gilbert writes in his book, *Stumbling on Happiness*, we often overestimate the magnitude of our unique responses to situations. Gilbert writes[92]:

> Our mythical belief in the variability and uniqueness of individuals is the main reason why we refuse to use others as surrogates. After all, surrogation is only useful when we can count on a surrogate to react to an event roughly as we would, and if we believe that people's emotional reactions are more varied than they actually are, then surrogation will seem less useful to us than it actually is. The irony, of course, is that surrogation is a cheap and effective way to predict one's future emotions, but because we don't realize how similar we all are, we reject this reliable method and rely instead on our imaginations, as flawed and fallible as they may be.

It seems that often a better predictor of how we respond to a situation is how others respond to that situation, rather than how we will imagine ourselves to do so based upon our current feelings and our current environment.

Does this apparent overestimation of our uniqueness carry over to physiologic, anatomic, behavioral, and health effects of nutritional, dietary, exercise, and other interventions? We do not know with

*Who actually first offered this quip is unclear: https://quoteinvestigator.com/2014/11/10/you-unique/.

certainty. In the sections above, we have offered methods to address these questions. Here, we broadly consider what we do and do not know about TRH to provide ideas for future research.

6.1 | To what extent is TRH in evidence?

That is, can we really be certain that treatment heterogeneity exists? In our opinion, for practical purposes, the answer seems to be yes. The work of Kaiser and Gadbury[1] in humans, work by Rikke et al. in mice,[93] the specific findings in the literature of observed moderators of treatment response,[94] and our everyday phenomenological experience all offer strong evidence that there is some TRH. This seems unequivocally true beyond any reasonable doubt.

6.2 | How large and common is the TRH?

Here, our knowledge is more meager, although interest is growing in devising approaches to estimate TRH variances in many scientific applications.[95,96] Yet, we do not have many carefully well-done analyses, across many situations, with rock-solid methods of estimation, in large samples, with good designs, and with good measurements to answer this unequivocally. Even in the rodent literature, some of the studies have been criticized for small sample sizes and other limits to experimental rigor, making their demonstrations of TRH seem dubious to some.[97] Thus, statements that it is "well known" that TRH variance is large relative to total variance in outcomes seem more an article of faith than a demonstrated fact. This point has been made in the specific context of the response to exercise training. A common statement in that context is that there is great variability in response. Yet, other statistically minded authors have used a model they believed allowed them to provide reasonable estimates of the variance in TRH in that domain and concluded that "evidence is limited for the notion that there are clinically important individual differences in exercise-mediated weight change."[7] The field may benefit from systematic sampling of the evidence space to better estimate the actual magnitude of TRH across many domains, treatments, types of people, and types of outcomes.

6.3 | Is the TRH that exists likely to be disordinal or ordinal?

By *ordinal* TRH, we mean situations in which the treatment that is better for some people is better for almost all people, but the degree to which it is better varies. In contrast, with *disordinal* TRH, we refer to circumstances in which the treatment that is better for some persons is actually poorer for others. Disordinal TRH is obviously more impactful than ordinal TRH because it fundamentally affects the choice of treatment. In contrast, with ordinal TRH, the degree of benefit will be better for some persons than for others, but the best treatment is still the same for everyone.

6.4 | To what extent is estimation of TRH in general, and specific contributors to it in particular, a feasible undertaking?

As is the paradox with every question about power analysis, the answer is always unknown. If one knew the exact magnitude of what one was looking for, one would have already found it and would not need to go looking. Yet, we can contemplate estimates. As more evidence emerges for the magnitude of TRH for different outcomes in different circumstances, we will begin to develop a sense of the magnitude of effects; the study designs that will best detect those effects; and the sample sizes, time periods, and expenses involved in conducting meaningful and powerful studies. In some simple cases, like evaluating whether effects differ by gender, or other common, easily observable, and roughly symmetrically distributed variables, detection should be relatively easy. In other cases, perhaps with complex genomic functions, this may not be the case.

By analogy, during the 1990s, statistical geneticists and genetic epidemiologists developed and refined the bulwark for detecting genes linked to (as opposed to directly associated with or causative of) complex phenotypes. The theory was well established, and it was clear that, in principle, we could detect such genes and had designs and analytic methods to do so. Yet, a barely kept secret was that, for practical purposes, almost all designs were so low in power that it was unclear whether, in humans, linkage analysis would ever be a useful tool for complex traits. Once genome-wide association studies became feasible, linkage analysis for complex traits in humans was almost completely abandoned. Genome-wide association studies, relying on far more powerful statistical underpinnings, began to find genes that could predict complex traits, in some cases likely identifying causal associations. However, such studies typically found associations that were so small that their practical utility in the clinical setting is dubious in most cases.

The findings of genome-wide association studies often point to interesting pathways and insights about traits[98] but seem not all that helpful in practical prediction or allocation of individuals to different treatments, at least in the obesity field.[99] Some people have proffered and claimed that they do have genomic or other related predictors of differential treatment response in the obesity domain,[100,101] but there does not seem to be large-scale, rigorous, replicated findings of clinically meaningful effects.[102] This invites the question of whether effects are likely to be not only scientifically demonstrable and replicable, but practically valid, that is, having practical utility. Importantly, when one considers practical utility, one must always ask, as Les McCann and the All-Stars did, "Compared to What?"

One comparison is compared with doing nothing. As stated above, if the TRH variance is small and/or strictly ordinal, a valid comparison is nothing with respect to differential treatment, or compared with giving everyone the same treatment. Alternatively, even if the TRH is larger and more meaningful, there may be simple ways of capitalizing on it from a clinically practical point of view that do not involve any deep understanding a priori. For example, one can use family history in place of more complex, genomic predictors or moderators of response. In turn, one can predict using physiologic moderators of response (e.g., basal insulin levels), to determine who will respond best to which treatment.[103] Doing so may be simpler than more complex genomic or other -omic approaches. Perhaps even simpler and more predictive still are elements of clinical trial-and-error. For example, it is common for people to recommend that an obesity pharmaceutical be given and if the patient does not lose a certain minimal degree of weight in an early time interval for the treatment be discontinued and something else tried instead. This seems to be a reasonable, practical approach to finding what works for the individual,[104] although it is not without the epistemological limitation of being unable to separate response from outcome. Thus, important future work will involve determining not only extensive TRH and the factors that underlie it, but its practical utility compared with more simple heuristics, and finding the right approach or combination of approaches for the specific setting and goals.

6.5 | What level of oversimplification is most useful?

Popper famously wrote "Science may be described as the art of systematic oversimplification—the art of discerning what we may with advantage omit."[105] We already use personalized medicine. We prescribe anti-hypertensive drugs to people with hypertension but not to people without hypertension. Yet, everyone would recognize this example as a rather crude level of personalization. Drilling down a level to say "Anybody with this polymorphism gets a different treatment or different dose of treatment than anybody without that polymorphism" is a far richer and more advanced form of personalized medicine or treatment provision. It seems likely that we will get to that point and in some cases already have.[106] But can we go further still to look at a single individual's unique combination of nuclear genomic nucleotides, microbiome, and metabolites and derive the prescription that is just right for them? While the answer, in theory, might be yes, in practice, this may be challenging. Each person is truly a unique combination of such bits of information, so how can one ever truly validate the results or estimate parts of a multivariate response when, in any data set, there is only one data point in that space? These are questions for the future, and we look forward to working together with others to address them and to watching the field unfold.

AFFILIATIONS

[1]Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, Indiana, USA
[2]Glanbia Performance Nutrition, Downers Grove, Illinois, USA
[3]Department of Psychology, University of South Carolina, Columbia, South Carolina, USA
[4]Department of Epidemiology and Biostatistics, Michigan State University, Lansing, Michigan, USA
[5]Biostatistics Program, School of Public Health, LSU Health Sciences Center, New Orleans, Louisiana, USA
[6]Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, Indiana, USA
[7]College of Population Health, Thomas Jefferson University, Philadelphia, Pennsylvania, USA
[8]Department of Genetics, Rutgers Robert Wood Johnson Medical School, New Brunswick, New Jersey, USA
[9]Department of Statistics, Kansas State University, Manhattan, Kansa, USA
[10]Department of Health Behavior, University of Alabama at Birmingham, Birmingham, Alabama, USA
[11]Departments of Epidemiology & Biostatistics and Statistics & Probability, IQ - Institute for Quantitative Health Science and Engineering, Michigan State University, Lansing, Michigan, USA
[12]Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, USA

ACKNOWLEDGMENTS

The authors acknowledge Jaya M. Satagopan for helpful comments and review.

CONFLICT OF INTEREST STATEMENT

BHE is now an employee of Glanbia Performance Nutrition, Downers Grove, IL, USA. However, this work was initiated while BHE was affiliated with the University of Illinois at Urbana-Champaign. DBA and his institution have received consulting fees, grants, contracts, and donations from multiple for-profit entities with interests in obesity, nutrition, statistics, and clinical trials, but none supported or are directly related to this manuscript. No other conflicts of interest are declared.

ORCID

Roger S. Zoh https://orcid.org/0000-0002-8066-1153
Bridget H. Esteves https://orcid.org/0000-0003-4588-9272
Amanda J. Fairchild https://orcid.org/0000-0001-8668-4658
Andrew G. Chapple https://orcid.org/0000-0001-5332-2730
Andrew W. Brown https://orcid.org/0000-0002-1758-8205
Luis M. Mestre https://orcid.org/0000-0001-7636-8133

REFERENCES

1. Kaiser KA, Gadbury GL. Estimating the range of obesity treatment response variability in humans: methods and illustrations. *Hum Hered*. 2013;75(2–4):127-135. doi:10.1159/000351738
2. Adams SH, Anthony JC, Carvajal R, et al. Perspective: guiding principles for the implementation of personalized nutrition approaches that benefit health and function. *Adv Nutr*. 2020;11(1):25-34. doi:10.1093/advances/nmz086
3. US Food and Drug Administration. *in Public Workshop on Patient-Focused Drug Development: Guidance 4—Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision Making*. Silver Spring; 2019.
4. Senn S. Mastering variation: variance components and personalised medicine. *Stat Med*. 2016;35(7):966-977. doi:10.1002/sim.6739
5. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701. doi:10.1037/h0037350
6. Senn S. Statistical pitfalls of personalized medicine. *Nature*. 2018;563(7733):619-621. doi:10.1038/d41586-018-07535-2

7. Williamson PJ, Atkinson G, Batterham AM. Inter-individual differences in weight change following exercise interventions: a systematic review and meta-analysis of randomized controlled trials. *Obes Rev*. 2018;19(7):960-975. doi:10.1111/obr.12682
8. Gadbury GL, Iyer HK. Unit–treatment interaction and its practical consequences. *Biometrics*. 2000;56(3):882-885. doi:10.1111/j.0006-341X.2000.00882.x
9. Poulson RS, Gadbury GL, Allison DB. Treatment heterogeneity and individual qualitative interaction. *Am Statistician*. 2012;66(1):16-24. doi:10.1080/00031305.2012.671724
10. Gadbury GL, Iyer HK, Allison DB. Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat*. 2001;11(4):313-333. doi:10.1081/BIP-120008851
11. Cortés J, González JA, Medina MN, et al. Does evidence support the high expectations placed in precision medicine? A bibliographic review. *F1000Research*. 2019;7:30. doi:10.12688/f1000research.13490.4
12. Senn S. Controversies concerning randomization and additivity in clinical trials. *Stat Med*. 2004;23(24):3729-3753. doi:10.1002/sim.2074
13. Senn S. Added values: controversies concerning randomization and additivity in clinical trials. *Stat Med*. 2004;23(24):3729-3753. doi:10.1002/sim.2074
14. Albert JM, Gadbury GL, Mascha EJ. Assessing treatment effect heterogeneity in clinical trials with blocked binary outcomes. *Biomet J: J Math Meth Biosci*. 2005;47(5):662-673. doi:10.1002/bimj.200510157
15. Gadbury GL, Iyer HK, Albert JM. Individual treatment effects in randomized trials with binary outcomes. *J Stat Planning Inference*. 2004;121(2):163-174. doi:10.1016/S0378-3758(03)00115-0
16. Holland PW. Statistics and causal inference. *J am Stat Assoc*. 1986;81(396):945-960. doi:10.1080/01621459.1986.10478354
17. Fairchild AJ, MacKinnon DP. Using mediation and moderation analyses to enhance prevention research. In: *Defining Prevention Science*. Springer; 2014:537-555. doi:10.1007/978-1-4899-7424-2_23
18. Fairchild AJ, McQuillin SD. Evaluating mediation and moderation effects in school psychology: a presentation of methods and review of current practice. *J Sch Psychol*. 2010;48(1):53-84. doi:10.1016/j.jsp.2009.09.001
19. Nunes EV, Pavlicova M, Hu MC, et al. Baseline matters: the importance of covariation for baseline severity in the analysis of clinical trials. *Am J Drug Alcohol Abuse*. 2011;37(5):446-452. doi:10.3109/00952990.2011.596980
20. Marre M, van Gaal L, Usadel KH, Ball M, Whatmough I, Guitard C. Nateglinide improves glycaemic control when added to metformin monotherapy: results of a randomized trial with type 2 diabetes patients. *Diabetes Obes Metab*. 2002;4(3):177-186. doi:10.1046/j.1463-1326.2002.00196.x
21. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *Jama*. 2006;296(10):1286-1289. doi:10.1001/jama.296.10.1286
22. Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Routledge; 2013.
23. Aiken LS, West SG, Reno RR. *Multiple Regression: Testing and Interpreting Interactions*. Sage; 1991.
24. VanderWeele T, Knol M. A tutorial on interaction. *Epidemiol Methods*. 2014;3:33-72.
25. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. *Int J Epidemiol*. 2012;41(2):514-520. doi:10.1093/ije/dyr218
26. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Vol. 3. Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
27. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol*. 1980;112(4):467-470. doi:10.1093/oxfordjournals.aje.a113015
28. Crona K. Rank orders and signed interactions in evolutionary biology. *Elife*. 2020;9:e51004. doi:10.7554/eLife.51004
29. Widaman KF, Helm JL, Castro-Schilo L, Pluess M, Stallings MC, Belsky J. Distinguishing ordinal and disordinal interactions. *Psychol Methods*. 2012;17(4):615-622. doi:10.1037/a0030003
30. Angus DC, Chang C-CH. Heterogeneity of treatment effect: estimating how the effects of interventions vary across individuals. *Jama*. 2021;326(22):2312-2313. doi:10.1001/jama.2021.20552
31. Brown AW, Bohan Brown MM, Onken KL, Beitz DC. Short-term consumption of sucralose, a nonnutritive sweetener, is similar to water with regard to select markers of hunger signaling and short-term glucose homeostasis in women. *Nutr Res*. 2011;31(12):882-888. doi:10.1016/j.nutres.2011.10.004
32. Renjilian DA, Perri MG, Nezu AM, McKelvey WF, Shermer RL, Anton SD. Individual versus group therapy for obesity: effects of matching participants to their treatment preferences. *J Consult Clin Psychol*. 2001;69(4):717-721. doi:10.1037/0022-006X.69.4.717
33. Gardner CD, Trepanowski JF, del Gobbo LC, et al. Effect of low-fat vs low-carbohydrate diet on 12-month weight loss in overweight adults and the association with genotype pattern or insulin secretion: the DIETFITS randomized clinical trial. *Jama*. 2018;319(7):667-679. doi:10.1001/jama.2018.0245
34. Brown AW, Hang J, Dussault PH, Carr TP. Plant sterol and stanol substrate specificity of pancreatic cholesterol esterase. *J Nutr Biochem*. 2010;21(8):736-740. doi:10.1016/j.jnutbio.2009.04.008
35. Brown AW, Hang J, Dussault PH, Carr TP. Phytosterol ester constituents affect micellar cholesterol solubility in model bile. *Lipids*. 2010;45(9):855-862. doi:10.1007/s11745-010-3456-6
36. Allen PS, Brown AW, Brown MM, Hsu WH, Beitz DC. Taurine and vitamin E supplementations have minimal effects on body composition, hepatic lipids, and blood hormone and metabolite concentrations in healthy Sprague Dawley rats. *Nutr Diet Suppl*. 2015;7:77-85.
37. McClelland GH, Judd CM. *Statistical Difficulties of Detecting Interactions and Moderator Effects*. American Psychological Association; 1993:376-390.
38. Shieh G. Sample size determination for examining interaction effects in factorial designs under variance heterogeneity. *Psychol Methods*. 2018;23(1):113-124. doi:10.1037/met0000150
39. Rouder JN, Engelhardt CR, McCabe S, Morey RD. Model comparison in ANOVA. *Psychon Bull Rev*. 2016;23(6):1779-1786. doi:10.3758/s13423-016-1026-5
40. Senn SS. *Cross-Over Trials in Clinical Research*. Vol. 5. John Wiley & Sons; 2002. doi:10.1002/0470854596
41. Gewandter JS, McDermott MP, He H, et al. Demonstrating heterogeneity of treatment effects among patients: an overlooked but important step toward precision medicine. *Clin Pharmacol Ther*. 2019;106(1):204-210. doi:10.1002/cpt.1372
42. Ezzet F, Whitehead J. A random effects model for binary data from crossover clinical trials. *J R Stat Soc Ser C Appl Stat*. 1992;41(1):117-126. doi:10.2307/2347622
43. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med*. 2009;28(2):221-239. doi:10.1002/sim.3478
44. McCulloch CE, Searle SR. *Generalized, Linear, and Mixed Models*. John Wiley & Sons; 2004.
45. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. In: Hall CA, ed. *Journal of the Royal Statistical Society Series A*. Vol.178. Chapman and Hall; 2004.
46. Schmid CH, Brown EN. Bayesian hierarchical models. In: *Methods in Enzymology*. Elsevier; 2000:305-330. doi:10.1016/S0076-6879(00)21200-7
47. Huber AM, Tomlinson GA, Koren G, Feldman BM. Amitriptyline to relieve pain in juvenile idiopathic arthritis: a pilot study using Bayesian metaanalysis of multiple N-of-1 clinical trials. *J Rheumatol*. 2007;34(5):1125-1132.

48. Mitchell GK, Hardy JR, Nikles CJ, et al. The effect of methylphenidate on fatigue in advanced cancer: an aggregated n-of-1 trial. *J Pain Symptom Manage*. 2015;50(3):289-296. doi:10.1016/j.jpainsymman.2015.03.009
49. Nathan P, Tomlinson G, Dupuis LL, et al. A pilot study of ondansetron plus metopimazine vs. ondansetron monotherapy in children receiving highly emetogenic chemotherapy: a Bayesian randomized serial N-of-1 trials design. *Support Care Cancer*. 2006;14(3):268-276. doi:10.1007/s00520-005-0875-7
50. Sung L, Tomlinson GA, Greenberg ML, et al. Serial controlled N-of-1 trials of topical vitamin E as prophylaxis for chemotherapy-induced oral mucositis in paediatric patients. *Eur J Cancer*. 2007;43(8):1269-1275. doi:10.1016/j.ejca.2007.02.001
51. Balaam L. A two-period design with t2 experimental units. *Biometrics*. 1968;24(1):61-73. doi:10.2307/2528460
52. Laird NM, Skinner J, Kenward M. An analysis of two-period crossover designs with carry-over effects. *Stat Med*. 1992;11(14–15):1967-1979. doi:10.1002/sim.4780111415
53. Candel M. Parallel, AA/BB, AB/BA and Balaam's design: efficient and maximin choices when testing the treatment effect in a mixed effects linear regression. *Pharm Stat*. 2012;11(2):97-106. doi:10.1002/pst.502
54. Mori J, Kano Y. A powerful test for Balaam's design. *Pharm Stat*. 2015;14(6):464-470. doi:10.1002/pst.1703
55. Loop MS, Frazier-Wood AC, Thomas AS, et al. Submitted for your consideration: potential advantages of a novel clinical trial design and initial patient reaction. *Front Genet*. 2012;3:145. doi:10.3389/fgene.2012.00145
56. Araujo A, Julious S, Senn S. Understanding variation in sets of N-of-1 trials. *PLoS ONE*. 2016;11(12):e0167167. doi:10.1371/journal.pone.0167167
57. Potter T, Vieira R, de Roos B. Perspective: application of N-of-1 methods in personalized nutrition research. *Adv Nutr*. 2021;12(3):579-589. doi:10.1093/advances/nmaa173
58. Friedman LM, Furberg CD, DeMets DL, et al. The randomization process. In: Friedman LM et al., eds. *Fundamentals of Clinical Trials*. Springer International Publishing; 2015:123-145. doi:10.1007/978-3-319-18539-2_6
59. Shadish WR. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin; 2001. doi:10.1016/B0-08-043076-7/00419-8
60. Brewin CR, Bradley C. Patient preferences and randomised clinical trials. BMJ. *Br Med J*. 1989;299(6694):313-315.
61. Long Q, Little RJ, Lin X. Causal inference in hybrid intervention trials involving treatment choice. *J am Stat Assoc*. 2008;103(482):474-484. doi:10.1198/016214507000000662
62. Marcus SM, Stuart EA, Wang P, Shadish WR, Steiner PM. Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychol Methods*. 2012;17(2):244-254. doi:10.1037/a0028031
63. King M, Nazareth I, Lampe F, et al. Impact of participant and physician intervention preferences on randomized trials: a systematic review. *Jama*. 2005;293(9):1089-1099. doi:10.1001/jama.293.9.1089
64. Torgerson DJ, Klaber-Moffett J, Russell IT. Patient preferences in randomised trials: threat or opportunity? *J Health Serv Res Policy*. 1996;1(4):194-197. doi:10.1177/135581969600100403
65. Zelen M. Randomized consent designs for clinical trials: an update. *Stat Med*. 1990;9(6):645-656. doi:10.1002/sim.4780090611
66. Torgerson DJ, Roland M. What is Zelen's design? *BMJ (Clin Res Ed)*. 1998;316(7131):606. doi:10.1136/bmj.316.7131.606
67. Heo M, Meissner P, Litwin AH, et al. Preference option randomized design (PORD) for comparative effectiveness research: statistical power for testing comparative effect, preference effect, selection effect, intent-to-treat effect, and overall effect. *Stat Methods Med Res*. 2017;28(2):626-640. doi:10.1177/0962280217734584
68. Rucker G. A two-stage trial design for testing treatment, self-selection and treatment preference effects. *Stat Med*. 1989;8(4):477-485. doi:10.1002/sim.4780080411
69. Preference Collaborative Review, G. Patients' preferences within randomised trials: systematic review and patient level meta-analysis. *BMJ (Clin Res Ed)*. 2008;337(oct31 1):a1864. doi:10.1136/bmj.a1864
70. Yancy WS Jr, Coffman CJ, Geiselman PJ, et al. Considering patient diet preference to optimize weight loss: design considerations of a randomized trial investigating the impact of choice. *Contemp Clin Trials*. 2013;35(1):106-116. doi:10.1016/j.cct.2013.03.002
71. Yancy WS Jr, Mayer SB, Coffman CJ, et al. Effect of allowing choice of diet on weight loss: a randomized trial. *Ann Intern Med*. 2015;162(12):805-814. doi:10.7326/M14-2358
72. Burke LE, Warziski M, Styn MA, Music E, Hudson AG, Sereika SM. A randomized clinical trial of a standard versus vegetarian diet for weight loss: the impact of treatment preference. *Int J Obes (Lond)*. 2008;32(1):166-176. doi:10.1038/sj.ijo.0803706
73. US Food and Drug Administration. *Guidance for Industry: Frequently Asked Questions About GRAS*. 2004 [cited 2020 August 12]; Available from: http://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/IngredientsAdditivesGRASPackaging/ucm061846.htm
74. Elliman NA, Ash J, Green MW. Pre-existent expectancy effects in the relationship between caffeine and performance. *Appetite*. 2010;55(2):355-358. doi:10.1016/j.appet.2010.03.016
75. Kelemen WL, Kaighobadi F. Expectancy and pharmacology influence the subjective effects of nicotine in a balanced-placebo design. *Exp Clin Psychopharmacol*. 2007;15(1):93-101. doi:10.1037/1064-1297.15.1.93
76. Waring DR. The antidepressant debate and the balanced placebo trial design: an ethical analysis. *Int J Law Psychiatry*. 2008;31(6):453-462. doi:10.1016/j.ijlp.2008.09.001
77. George BJ, Li P, Lieberman HR, et al. Randomization to randomization probability: estimating treatment effects under actual conditions of use. *Psychol Methods*. 2018;23(2):337-350. doi:10.1037/met0000138
78. Eichelbaum M, Spannbrucker N, Steincke B, Dengler HJ. Defective N-oxidation of sparteine in man: a new pharmacogenetic defect. *Eur J Clin Pharmacol*. 1979;16(3):183-187. doi:10.1007/BF00562059
79. Mahgoub A, Dring LG, Idle JR, Lancaster R, Smith RL. Polymorphic hydroxylation of debrisoquine in man. *Lancet*. 1977;310(8038):584-586. doi:10.1016/S0140-6736(77)91430-1
80. McDonnell AM, Dang CH. Basic review of the cytochrome p450 system. *J Adv Pract Oncol*. 2013;4(4):263-268. doi:10.6004/jadpro.2013.4.4.7
81. Jurado-Escobar R, Perkins JR, García-Martín E, et al. Update on the genetic basis of drug hypersensitivity reactions. *J Investig Allergol Clin Immunol*. 2017;27(6):336-345. doi:10.18176/jiaci.0199
82. Prat A, Ellis MJ, Perou CM. Practical implications of gene-expression-based assays for breast oncologists. *Nat Rev Clin Oncol*. 2012;9(1):48-57. doi:10.1038/nrclinonc.2011.178
83. Sørlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci*. 2003;100(14):8418-8423. doi:10.1073/pnas.0932692100
84. Liu MC, Pitcher BN, Mardis ER, et al. PAM50 gene signatures and breast cancer prognosis with adjuvant anthracycline-and taxane-based chemotherapy: correlative analysis of C9741 (Alliance). *NPJ Breast Cancer*. 2016;2(1):15023. doi:10.1038/npjbcancer.2015.23
85. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120

86. Giacomini KM, Yee SW, Mushiroda T, Weinshilboum RM, Ratain MJ, Kubo M. Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nat Rev Drug Discov*. 2017;16(1):70. doi:10.1038/nrd.2016.234
87. MacLean PS, Rothman AJ, Nicastro HL, et al. The accumulating data to optimally predict obesity treatment (ADOPT) core measures project: rationale and approach. *Obesity*. 2018;26(Suppl 2):S6-s15. doi:10.1002/oby.22154
88. Bray MS, Loos RJF, McCaffery JM, et al. NIH working group report-using genomic information to guide weight management: from universal to precision treatment. *Obesity*. 2016;24(1):14-22. doi:10.1002/oby.21381
89. Scott RA. Unraveling the role for genetics in enabling precision prescription for weight loss-scaling up for success. *Obesity*. 2016;24(1):12-13. doi:10.1002/oby.21380
90. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581-590. doi:10.1038/s41576-018-0018-x
91. de los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat Rev Genet*. 2010;11(12):880-886. doi:10.1038/nrg2898
92. Gilbert D, Lyngdoh T. *Stumbling on Happiness*. SAGE Publications Sage India; 2015.
93. Rikke BA, Battaglia ME, Allison DB, Johnson TE. Murine weight loss exhibits significant genetic variation during dietary restriction. *Physiol Genomics*. 2006;27(2):122-130. doi:10.1152/physiolgenomics.00068.2006
94. Bomberg EM, Ryder JR, Brundage RC, et al. Precision medicine in adult and pediatric obesity: a clinical perspective. *Ther Adv Endocrinol Metab*. 2019;10:2042018819863022.
95. Beavers DP, Hsieh KL, Kitzman DW, et al. Estimating heterogeneity of physical function treatment response to caloric restriction among older adults with obesity. *PLoS ONE*. 2022;17(5):e0267779. doi:10.1371/journal.pone.0267779
96. Volkmann C, Volkmann A, Müller CA. On the treatment effect heterogeneity of antidepressants in major depression: a Bayesian meta-analysis and simulation study. *PLoS ONE*. 2020;15(11):e0241497. doi:10.1371/journal.pone.0241497
97. Mattson MP. Genes and behavior interact to determine mortality in mice when food is scarce and competition fierce. *Aging Cell*. 2010;9(3):448-449. doi:10.1111/j.1474-9726.2010.00561.x
98. Speakman JR, Loos RJF, O'Rahilly S, Hirschhorn JN, Allison DB. GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. *Int J Obes (Lond)*. 2018;42(8):1524-1531. doi:10.1038/s41366-018-0147-5
99. El-Sayed Moustafa JS, Froguel P. From obesity genetics to the future of personalized obesity therapy. Nature reviews. *Endocrinology*. 2013;9(7):402-413. doi:10.1038/nrendo.2013.57
100. Wang J, García-Bailo B, Nielsen DE, el-Sohemy A. ABO genotype, 'blood-type' diet and cardiometabolic risk factors. *PLoS ONE*. 2014;9(1):e84749. doi:10.1371/journal.pone.0084749
101. Zhang X, Qi Q, Zhang C, et al. FTO genotype and 2-year change in body composition and fat distribution in response to weight-loss diets: the POUNDS LOST trial. *Diabetes*. 2012;61(11):3005-3011. doi:10.2337/db11-1799
102. Bayer S, Winkler V, Hauner H, Holzapfel C. Associations between genotype-diet interactions and weight loss—a systematic review. *Nutrients*. 2020;12(9):2891. doi:10.3390/nu12092891
103. Wong JM, Yu S, Ma C, et al. Stimulated insulin secretion predicts changes in body composition following weight loss in adults with high BMI. *J Nutr*. 2022;152(3):655-662. doi:10.1093/jn/nxab315
104. Cefalu WT, Bray GA, Home PD, et al. Advances in the science, treatment, and prevention of the disease of obesity: reflections from an editors' expert forum. *Diabetes Care*. 2015;38(8):1567-1582. doi:10.2337/dc15-1081
105. Popper, K., *The Open Universe*. 1982.
106. Goulding R, Dawes D, Price M, Wilkie S, Dawes M. Genotype-guided drug prescribing: a systematic review and meta-analysis of randomized control trials. *Br J Clin Pharmacol*. 2015;80(4):868-877. doi:10.1111/bcp.12475

How to cite this article: Zoh RS, Esteves BH, Yu X, et al. Design, analysis, and interpretation of treatment response heterogeneity in personalized nutrition and obesity treatment research. *Obesity Reviews*. 2023;24(12):e13635. doi:10.1111/obr.13635