Published in final edited form as:

Hum Hered. 2013 ; 75(0): 127–135. doi:10.1159/000351738.

Estimating the Range of Obesity Treatment Response Variability in Humans: Methods and Illustrations

Kathryn A. Kaiser, Ph.D.¹ and Gary L. Gadbury, Ph.D.²

¹Office of Energetics, Dean's Office, School of Public Health, University of Alabama at Birmingham, Alabama, U.S.A.

²Department of Statistics, Kansas State University, Manhattan, Kansas, U.S.A.

Abstract

Background/Aims—The rising prevalence of human obesity worldwide has focused research on a variety of interventions that result in highly varied degrees of weight loss (WL). The advent of genomic testing has quantified estimates of both the contribution of genetic factors to the development of obesity as well as racial/ethnic variation of risk alleles across sub-populations. More recent studies have examined genetic associations with effectiveness of WL interventions, but to date are unable to explain a large proportion of the variance observed.

Methods—We describe and provide two illustrations of statistical methods to estimate upper and lower bounds of WL treatment response heterogeneity (TRH) in the absence of genotypic data, using published summary statistics and a raw dataset from weight loss studies.

Results—Thirty-two studies had some evidence of a positive mean treatment effect with respect to the control intervention. Twelve of these 32 studies exhibited WL TRH. Of these 12, three exhibited an estimated proportion of >5% of the sampled population having an outcome opposite the mean effect. In the raw dataset, bounds estimations for change in waist circumference revealed tighter ranges in men than women.

Conclusion—Future studies may be able to take advantage of multiple approaches, including the method we describe, to identify and quantify the presence of TRH in studies of WL or related outcomes.

Keywords

obesity; diet; exercise; human variability; treatment heterogeneity; tightening bounds

INTRODUCTION

Examinations in the scientific literature have evaluated genetic associations between specific genes, loci and single-nucleotide polymorphisms and human and/or animal obesity. The heritability estimate of human body mass index (BMI) is approximately 0.65 [1;2].

Additionally, some genetic associations have been examined pertaining to human and animal thinness [3]. More recently, studies have been performed to evaluate the genetic influence on the weight loss (WL) process, seeking to understand if the use of genetic information about a WL patient can determine a more personalized treatment plan or predict the amount of weight that can be lost [4;5]. Yet to date, genetic associations that disappointingly explain only a small amount (< 7%) of the variance in BMI [6] within

populations and similarly small magnitudes of weight loss being explained by genotypes are reported in WL studies [5]. Our focus here is on variability in treatment response across individuals, such as the effect of an intervention on BMI loss versus a control intervention. This is what is referred to herein as WL treatment response variability or heterogeneity.

Of the current methods for attempting WL, bariatric surgery has the largest average effect (loss from baseline weight ranging from 15-30% within the first year post-surgery) [7]. Yet there is significant weight regain within five years after the surgery, although some patients maintain WL for up to ten years [8]. Outcomes (i.e., gains or losses that are observed over time) are highly variable, depending upon other aspects of post-surgical care [9]. Of the more common WL methods using dietary and physical activity interventions, outcomes average 2-5 Kg with diet alone and about 0-2 Kg with exercise alone [10]. These effects often plateau within about six months with some regain typically seen by 12 months [10]. Observed WL effects often exhibit non-normal distributions with large variances [4]. Thus far, sufficient predictive models using genetic information are not available to reliably predict if an individual will respond to a given WL intervention and to what degree [4]. It is often poorly understood how much treatment response heterogeneity might be present in the first place and whether a search for such gene-treatment interactions that explain this heterogeneity will be fruitful [11]. However, much data exists to estimate the range of likely responses to present day treatments. Such analytic approaches can inform future investigations of the potential genetic contribution of variability in WL treatment response in humans.

Present aim

We aimed to quantify the upper and lower bounds of WL treatment response heterogeneity using summary statistics from published randomized controlled trials and a raw data set. First, we discuss the statistical considerations for such approaches, followed by two example illustrations using summary statistics from 66 published trials and raw data from a single, large trial.

VARIABILITY IN TREATMENT RESPONSE

Variability in treatment response refers to the idea that the true effect of a treatment for one individual will be different from that of another. Some have called this “individual treatment heterogeneity” [12] or “subject-treatment interaction” [13]. (N.B. - Throughout this discussion, we refer only to statistical interactions and do not attempt to quantify specific gene interactions – see [14] for a more thorough explanation of types of interactions within the genetic framework). To formally define this idea, suppose an outcome of interest (e.g., change from baseline in weight or BMI) after exposure to a treatment, noted as b , is given by Y_b , and that the outcome after exposure to a different treatment, a (perhaps a control treatment) is given by Y_a . The bivariate pair of outcomes, (Y_a, Y_b) , are called potential outcomes [15] and the potential outcomes framework has become widely used in statistics for studying causal effects in randomized and non-randomized settings.

A true, singular, causal effect of treatment b with respect to treatment a , at a particular point in time, is given by $D = Y_b - Y_a$, which represents an individual WL of a subject when on treatment b versus treatment a (control). D cannot be observed for an individual since an individual is either assigned to treatment or control at a given time. This has been referred to as the fundamental problem of causal inference [16]. It is worth noting that a change in a particular subject's weight over a given time period may be due to the effect of treatment or other factors (observed or unobserved) that were at play during the time period. The true effect of the treatment for an individual cannot be separated from these other factors.

Often, $E(D) = \mu_D$ (mean treatment effect) is of greatest interest. Here, we consider a measure of individual treatment heterogeneity given by the standard deviation, $\sigma_D = \sqrt{\text{Var}(D)}$. This standard deviation cannot be estimated from observable data because there is no information on the correlation between Y_b and Y_a . Yet, estimable bounds exist that are defined by setting this non-estimable correlation equal to 1 and -1 . We denote these bounds as $\sigma_D^{\min}, \sigma_D^{\max}$, where

$$\sigma_D^{\min} = |\sigma_b - \sigma_a| \text{ and } \sigma_D^{\max} = (\sigma_b + \sigma_a), \quad (1)$$

and where σ_b, σ_a are the standard deviations of outcomes on treatments b and a , respectively. Evidence of heteroscedasticity across the two treatment groups is an indication that treatment heterogeneity is present in the population from which samples were obtained. If outcomes are assumed to be normally distributed, then a usual F-test for heteroscedasticity, i.e., a test for $\sigma_b = \sigma_a$, is also a test for whether $\sigma_D^{\min} > 0$. This particular test can be conducted based on summary statistics from a study or several studies, again assuming that outcomes are at least approximately normally distributed. We illustrate this in the next section. When treatment heterogeneity is likely to be present in the population from which samples were obtained, a secondary question is whether this possible heterogeneity has important consequences that should be investigated. One consequence of this heterogeneity is that the effect of a treatment b with respect to a may be in opposite directions across different individuals or subsets of individuals, with one treatment appearing to have higher efficacy for some and the other appearing to have higher efficacy for others. The term qualitative interaction (QI) has been used to describe this situation at the subset level [17;18]. A “quantitative” interaction [18] exists when the magnitudes of the effect of treatment b versus treatment a differ across subsets, but are in the same direction. Transformations are sometimes used to remove interactions so that, on the transformed scale, more comprehensive statements about treatment effects can be made. If measurements are obtained on a clinically meaningful scale and subject-treatment interaction is present in the population on that scale, transforming the data with monotonic but non-linear transformations may eliminate such interactions. Doing so is not inherently correct or incorrect but, with clinically meaningful scales, one might argue that any assessment of subject-treatment interaction should be done on the original scale of measurement. However, if there are no qualitative interactions present in the population when using a clinical meaningful scale of measurement, then for some applications the question of variability is of lesser consequence because one treatment appears to be superior to another across all subjects. In such a case, a transformation may restore homoscedasticity and facilitate the analysis.

In studies that exhibit evidence of a positive mean treatment effect, $\mu_D > 0$, treatment heterogeneity could result in a proportion of individuals that have an effect in the opposite direction from the mean, i.e., a value of $D < 0$. If an outcome such as WL is assumed to be a

variable that is normally distributed, then this proportion is quantified by $P_{-D} = \Phi\left(\frac{-\mu_D}{\sigma_D}\right)$ where Φ denotes the cumulative distribution function of a standard normal distribution. P_{-D} cannot be estimated for the same reason that σ_D cannot be estimated. However, estimable bounds given by,

$$(P_{-D}^{\min}, P_{-D}^{\max}) = \left(\Phi\left(\frac{-\mu_D}{\sigma_D^{\min}}\right), \Phi\left(\frac{-\mu_D}{\sigma_D^{\max}}\right) \right). \quad (2)$$

The role of a covariate

The above discussion does not consider the use of information in other variables that is available in most human WL studies. Here we consider the role of commonly used covariates – those variables not affected by treatment, such as age, gender, race, and other demographics. We also examine effects of treatment using baseline body composition in the model. These variables play two roles: 1) they may provide further evidence that treatment heterogeneity is present and 2) they allow us to refine or tighten estimable bounds for treatment heterogeneity.

Consider, as in [19], a continuous covariate Z with mean μ_Z and variance σ_Z^2 that augments the two potential outcome variables (Y_a, Y_b) . Let $\beta_{Y_b|Z}$ and $\beta_{Y_a|Z}$ be the population regression coefficients relating Z to Y_b and to Y_a , respectively. Let $\sigma_{Y_b|Z}$ and $\sigma_{Y_a|Z}$ be the standard deviations of the conditional distributions of Y_b and Y_a , respectively, given Z , and let $\rho_{Y_b Y_a|Z}$ be the partial correlation between Y_b and Y_a given Z . Then it has been shown [19] that,

$$\sigma_D^2 = \left(\sigma_{Y_b|Z} - \sigma_{Y_a|Z} \right)^2 + \left(\beta_{Y_b|Z} - \beta_{Y_a|Z} \right)^2 \sigma_Z^2 + 2 \sigma_{Y_b|Z} \sigma_{Y_a|Z} \left(1 - \rho_{Y_b Y_a|Z} \right). \quad (3)$$

Several observations can be made regarding this expression. First, all parameters in it can be estimated using observed data except the partial correlation. Second, if Z is a perfect linear predictor of Y_b and of Y_a , then all subject-treatment (S-T) interaction is explained by a covariate-treatment interaction since $\sigma_{Y_b|Z} = \sigma_{Y_a|Z} = 0$. Third, if this is not the case, then $\sigma_D^2 = 0$ if and only if $\sigma_{Y_b|Z} = \sigma_{Y_a|Z}$, $\beta_{Y_b|Z} = \beta_{Y_a|Z}$, and $\rho_{Y_b Y_a|Z} = 1$, and the last equality cannot be tested using observed data. Thus, either the presence of heteroscedasticity of residual variances across the two treatment groups or a covariate-treatment interaction (or both) provides evidence that treatment heterogeneity is present. A similar result holds for a categorical covariate and was shown in [12]. With a continuous covariate, tightened bounds

for σ_D become, $\sigma_D^{\min} = \sqrt{\left(\sigma_{Y_b|Z} - \sigma_{Y_a|Z} \right)^2 + \left(\beta_{Y_b|Z} - \beta_{Y_a|Z} \right)^2 \sigma_Z^2}$ and

$\sigma_D^{\max} = \sqrt{\left(\sigma_{Y_b|Z} + \sigma_{Y_a|Z} \right)^2 + \left(\beta_{Y_b|Z} - \beta_{Y_a|Z} \right)^2 \sigma_Z^2}$. The term $\left(\beta_{Y_b|Z} - \beta_{Y_a|Z} \right)^2 \sigma_Z^2$ represents the explainable portion of treatment heterogeneity – explainable by the use of the covariate, Z . The unexplainable portion of treatment heterogeneity has estimable bounds resembling those given in equation (1) but using conditional standard deviations, conditioned on Z . These bounds can be interpreted as bounds for treatment heterogeneity of subpopulations defined by values of Z , that is, bounds for $\sigma_{D|Z}$. If the mean treatment effect is positive for a given sub-population, then the proportion of the sub-population having a negative effect are bounded by quantities like those in equation (2), but instead, using the conditional mean and standard deviations of sub-populations given a value of Z . An example of this is given in [12].

Results shown above for the use of one covariate generalize to the use of multiple covariates, and the tightness of the bounds depend on the predictive capability of the set of covariates, that is, their ability to reduce the conditional standard deviations $\sigma_{Y_b|Z}$ and $\sigma_{Y_a|Z}$. The derivation of the technique for using multiple covariates is given in the appendix.

RESULTS

Evaluating treatment heterogeneity in summary statistics from multiple studies

In a literature search of adult obesity randomized controlled trials (RCTs) published between January 2007 and July 2009, 166 studies were identified that reported measurements of

weight change, BMI change, or both. Of these, 66 studies were identified that reported mean and standard deviations for WL as well as mean and standard deviations for baseline weight (see bibliography of included studies in Supplemental Material). All studies were evaluated according to the following inclusion criteria: 1) the data were from human studies, 2) the study was an RCT, 3) the study had a total sample size of at least 30 participants at enrollment, 4) the study protocol included an intervention period of at least 8 weeks, 5) weight loss and/or weight gain prevention was a primary or secondary outcome variable, 6) the publication was available in the English language. Some of these studies had multiple treatment arms but the present focus is only the primary intervention arm and the control arm. Of the 66 studies initially analyzed, intervention types were: 25.8% behavioral/lifestyle/educational, 36.4% diet only, 4.5% exercise only, 21.2% mixed (more than one of any type), and 12.1% pharmaceutical. For these 66 studies, single or double blind designs were reported in 51.5% overall, and 17.6% behavioral/lifestyle/educational, 70.8% diet, 33.3% exercise, 35.7% mixed, and 100% pharmaceutical studies reporting single or double blind designs. The intervention periods ranged from 8 weeks to 3 years, with a mean of 29 weeks and a mode of 12 weeks.

To assess what proportion of studies exhibit evidence of treatment heterogeneity, a

hypothesis test of the form $H_0: \frac{\sigma_a^2}{\sigma_b^2} = 1$ was evaluated against a two-tailed alternative, where σ_a^2 and σ_b^2 are the population variances of WL in control and treatment groups, respectively.

This was used as a test for evaluating whether there was evidence that $\sigma_D^{\min} > 0$. Of the 66 studies, 17 (26%) showed evidence that treatment heterogeneity was present, as compared to three (5%) that would have been expected by random chance. Thus, one could surmise that an estimated proportion of 0.26 (with standard error = 0.05) studies exhibit evidence of treatment heterogeneity, solely on the basis of summary measures that were available on the studies considered here. There is no clear evidence of association between those studies that showed evidence of treatment heterogeneity and the primary intervention type, the duration of the intervention, or the baseline mean weight/BMI.

We now illustrate the use of estimated values of $(P_{-D}^{\min}, P_{-D}^{\max})$, given by equation (2). Of the 66 studies, 32 indicated a positive mean treatment effect that was at least marginally significant (a standardized mean difference greater than 2). Of these 32, 12 exhibited evidence that treatment heterogeneity was present, i.e., $\sigma_D^{\min} > 0$. However, only 3 studies had a point estimate of $P_{-D}^{\min}$ that exceeded 0.05 (i.e., an estimated minimum percentage of the population having a negative treatment effect greater than 5%). This suggests that a proportion of 0.094 (i.e., 3 out of 32) of sampled studies indicate treatment heterogeneity where more than 5% of a population with a treatment effect in the opposite direction from the mean effect. An effective method to estimate standard errors for estimated $P_{-D}^{\min}$ is the use of bootstrap techniques as was done in [19]. However, this requires raw data that were not available for all of these studies.

Of the 32 studies with evidence of a positive mean treatment effect, four had an estimated $P_{-D}^{\max}$ that was less than 0.09 (the estimated $P_{-D}^{\min}$ was close to zero for these four). These studies had a relatively large estimated mean effect and a small estimate for $\sigma_D^{\max}$. When the estimated $P_{-D}^{\max}$ is small, it suggests that most of the population of individuals should have an effect of treatment in agreement with the mean effect. For three of these four studies, diet was the primary intervention. All bounds were estimated using only information in summary statistics on the primary outcome variable, in this case WL. Bounds can be tightened using covariate information, but access to the raw data is needed or substantially more information

in summary statistics regarding the quantitative relationships between covariates and outcomes. Next, we investigate this approach to tightening bounds of WL treatment response.

Illustration with a raw data set and covariates

The PREMIER weight loss trial [20] was a randomized trial to investigate the effects of two multi-component lifestyle interventions on blood pressure in a large sample of 761 adults (61.9% female, 34% African American) who were assigned to either : 1) a behavioral lifestyle (BLS) intervention that implemented established recommendations, 2) a BLS intervention of established recommendations plus the DASH (Dietary Approaches to Stop Hypertension) diet, or 3) an advice-only standard of care group for a period of 18 months. Participants were ($M \pm SD$) 49.3 ± 8.8 years old and had a BMI of 33.0 ± 5.7 Kg/m² at study entry [4;21].

Here we illustrate treatment heterogeneity using waist circumference (WC) loss as a primary outcome (positive numbers were a loss in WC during the trial). We selected this outcome variable as it has predictive utility in health outcomes and is linked to genetic influences [4;21]. The treatment group considered is a “comprehensive plus DASH lifestyle intervention,” in which participants receive a behavioral intervention program designed to promote the DASH dietary pattern (increased intake of fruits, vegetables, and low-fat dairy products, and reduced intake of saturated fat and total fat) in addition to the longstanding recommendations for BP control (reduced salt intake, increased physical activity, reduced alcohol intake, and WL, if overweight). The control group is an “advice-only” control arm, in which participants receive information on how to reduce salt intake, increase physical activity, reduce alcohol intake, and lose weight if overweight. The information provided to participants is similar to that in the information-oriented programs that are sometimes provided as part of routine medical care.

The 253 individuals in the control arm lost an average WC of 1.17 cm with a standard deviation equal to 7.24 cm. The 254 subjects in the treatment arm lost an average WC of 3.64 cm with a standard deviation equal to 6.89 cm. The standardized mean difference is equal to 3.93 and a two-tailed t -test indicates that the treatment arm lost significantly more WC than the control arm ($p = 0.0001$).

Since the raw data are available, the bootstrap [22] can be used to estimate standard errors (SE) for estimated bounds for treatment heterogeneity. Lower and upper bounds for σ_D given in equation (1) are estimated to be 0.36 (SE=0.44) and 14.13 (SE=0.63), respectively. From the standard errors one can see that the lower bound is not significantly different from zero. An F-test for evaluating whether $\sigma_D^{\min} > 0$ is also not significant ($F=1.11$, $p=0.41$, two-tailed). However, only using the primary outcome data, σ_D is estimated to be as high as 14.13. The proportion of the population that could have lost more WC on the control arm versus the treatment arm is given by the bounds in equation (2). The estimated lower bound is very near zero, but the estimated upper bound is 0.43 with bootstrap standard error equal to 0.018. Without more information, one cannot tell whether it is more likely that all subjects would lose more WC on the treatment relative to control or whether up to 43% would have lost more WC by being assigned to the control condition instead.

We illustrate the use of baseline WC in explaining part of treatment heterogeneity when interactions may be present. This variable was chosen as the variable Z in the above discussion to illustrate the interaction effect shown in equation (3) for treatment heterogeneity. A model fitting WC loss against treatment and baseline WC shows a significant treatment-baseline interaction ($p = 0.01$). The estimated value of

$(\beta_{Y_bZ} - \beta_{Y_aZ})^2 \sigma_Z^2$ is equal to 2.45, which is the estimated amount of the variance of individual treatment effects, σ_D^2 , that is explained by the interaction. Refined bounds for σ_D are then estimated to be 1.56 (SE=0.57) and 13.56 (SE=0.59). The lower bound is now significantly different from zero, but the treatment heterogeneity is explained by an interaction between treatment and baseline WC. Further examination of the interaction reveals that the treatment has a significant effect on reducing WC relative to control for those individuals with a baseline WC below the third quartile. As baseline WC becomes larger, the treatment does not appear to have an effect that is different from the control condition. There is no evidence that the treatment is more effective than the control condition in reducing WC for these individuals above the third quartile at baseline.

Using several covariates, e.g., baseline values for weight, BMI, and WC as well as age, height, race, and sex, and using the derivation shown in the appendix, the estimated lower and upper bounds for σ_D are further tightened to be 3.03 and 12.92, respectively. Fitting a linear model with these covariates and all treatment covariate interactions, and then reducing it using backward stepwise selection using the AIC statistic as a criterion leads to a model with an R-square = 0.22. Thus, only 22% of the variability in WC loss is explained by the use of several covariates, indicating that much of the variability in WC change remains unexplained. Thus, unexplainable individual treatment heterogeneity remains. Tightening bounds further requires identification of variables that are predictive of success to one particular treatment versus another.

It may be of interest to evaluate whether one subset of a population exhibits more heterogeneity in treatment response than another. Subsets may be delineated by race, age categories, gender, etc. Here we illustrate such an evaluation using gender. Using no covariates, Table 1 shows results for a mean treatment effect and estimated $\sigma_D^{\min}$, $\sigma_D^{\max}$ and $(P_{-D}^{\min}, P_{-D}^{\max})$ for each gender separately. Standard errors were estimated using the nonparametric bootstrap [22] with 2000 samples within treatment groups. The results show that female subjects on treatment lost an average of 1.43 more cm of WC versus the control group. This difference was not statistically significant ($p = 0.107$). Male subjects on treatment lost 4.07 more cm than control with a p -value rounded to zero out to four decimal places. The bounds for treatment heterogeneity are tighter for male subjects versus females. There is an estimated minimum proportion of 0.065 of male subjects who would have a larger loss in WC on the control treatment, but the standard error shows this proportion to not be statistically different from zero. However, with no additional information, this proportion could be as large as 0.35. The estimated minimum proportion of female subjects who would lose more WC on the control condition is 0.19. The large standard error for this proportion is due to the mean effect not being statistically significant and the estimated minimum proportion can vary widely in such cases.

Table 2 shows the bounds for treatment heterogeneity using baseline WC as a covariate. One can see that the estimated bounds for both genders are tightened but more so for females. This is because the baseline WC and treatment-baseline interaction were highly significant for females ($p=0.001$) but not for males ($p=0.787$) (Figures 1A and 1B, respectively). Thus, the baseline covariate did not explain any treatment heterogeneity for male subjects, nor did it tighten the estimated bounds. Standard errors suggest that the estimated minimum bounds for both genders are greater than zero.

Table 3 shows tightened bounds using baseline weight, baseline WC, baseline BMI, age, and race as covariates. Using the formula in the appendix, one can see a further tightening of the bounds (particularly in males) as these additional covariates are added. Unlike baseline WC,

the added variables are significant in explaining variability in WC change due to treatment, thereby tightening estimated bounds for treatment heterogeneity. However, estimated minimum and maximum bounds remain wide in this model. Performing some variable selection in a linear model that includes interactions as covariates to select a ‘best’ model for each gender, the resulting R-square is only 0.22 for females and 0.21 for males, indicating that these variables only account for a relatively small proportion of the variability in WC change observed during the various interventions studied.

DISCUSSION AND SUMMARY

Our method and examples illustrate the application of determining bounds of treatment variability in human WL studies using existing data. Our estimates of upper bounds in the examples given show that as many as 30-40% of a sampled population could have had an effect of treatment in the opposite direction from the mean treatment effect. However, minimum bounds did not indicate that such a positive proportion must be present. Estimated bounds tended to be wide when only primary outcome variables were available. It is interesting to note that the F-test that is sometimes used for equality of variance in two sample studies also served to evaluate whether there is evidence of treatment heterogeneity in the population, i.e., $\sigma_D^{\min} > 0$.

In the example of the raw dataset, we illustrate a gender by treatment interaction. This type of interaction may be frequently overlooked in studies that do not analyze outcome data separately by gender for mixed samples. In turn, this could result in a reduced overall effect due to the choice of analysis and not the type of intervention. Future studies of genetic relationships to weight loss outcomes should examine gender data separately. The strengths of these analyses are in the rich datasets used for illustrating the methods described.

We must acknowledge the limitations of the conclusions drawn. First, replication with other datasets is advised as our selections were not comprehensive or systematic. The 66 studies used were a short date range (January 2007 – July 2009), but were part of a systematic review project. Second, the analysis and examples we discussed are unable to directly address quantification or specific interactions with potential genetic or epigenetic influences on WL treatment response. Additionally, an assumption of normal distributions of WL outcomes may not be supported in all studies. Further, we cannot quantitatively assess influence of bias due to non-blinded study designs on the heterogeneity of treatment response observed. Of the 66 studies initially examined, 51.5% had designs of either single or double blinding. Of the 32 assessed that demonstrated a positive mean treatment effect that was at least marginally significant, 48.4% had designs of either single or double blinding. The relative proportions of studies that have some type of blinding among the intervention types were also roughly equivalent in the 66 initial studies examined as compared to the 32 analyzed. For the pharmaceutical studies analyzed, we cannot separate the heterogeneity of response observed to any standard of care component in addition to the response to drug effects. Since these represented approximately 16% of the total studies analyzed, we estimate the influence of this type of potentially “additive” heterogeneity to be relatively small, but is not estimated in the present analysis. We provide a framework for future studies in which to add genetic or gene regulation data, allowing for quantification of the range of responses to WL interventions when individual risk genes (and/or gene regulation mechanisms) are more clearly identified as they pertain to the degree and variation of treatment response.

Estimated minimum and maximum bounds for treatment heterogeneity were computed by setting a non-estimable correlation (or partial correlation when covariates were used) equal to 1 and –1. In the case of no covariates, one might assume that the correlation between

potential outcomes (Y_a, Y_b) is more likely to be positive than negative because the two outcomes are defined to be potentially measurable on the same subject. Thus, bounds could be subjectively tightened by restricting the range of the non-estimable correlation to positive values. With covariates, one may assume that the two potential outcomes, given the covariates, may be in a restricted range (perhaps a range around zero) and, thus, facilitate tightening bounds on treatment heterogeneity. While these assumptions may certainly be plausible for certain applications, they are still unable to be verified with the observed data.

Other studies are necessary to further refine bounds for the unaccounted for variance, likely due to non-genetic or epigenetic components such as life-course factors (e.g., developmental factors, lifestyle habits, physiological adaptations to environments), unknown medical conditions, effects from receiving other WL treatments, and levels of compliance with treatments. To reduce these sources of variance, large and highly controlled studies (e.g., in metabolic units rather than in free-living samples) are needed. In the mean time, meta-analytic approaches can be applied to these areas where such existing data have been quantified to begin to refine estimates of the inter-individual variability of these components on weight loss outcomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the following persons for assistance in data acquisition and collection – Jack Smith, Greg Tarrant, Olivia Affuso, Tiffany Carson, Katherine Ingram, and Firas Abbas. We offer additional thanks to Andrew Brown and the anonymous reviewers for comments to improve the paper.

Funding Support: K. Kaiser – supported in part by R01DK078826, P30DK056336;

APPENDIX

Tightening bounds for σ_D with multiple covariates

Note that the variance of individual treatment effects is, $\sigma_D^2 = \sigma_{Y_a}^2 + \sigma_{Y_b}^2 - 2\sigma_{Y_a}\sigma_{Y_b}\rho_{Y_bY_a}$, where the bounds for individual treatment heterogeneity, using no covariates, were obtained by setting the non-estimable correlation between the two potential outcome variables, $\rho_{Y_bY_a}$, equal to 1 and -1. Therefore, tightening the bounds depends on tightening the bounds for this correlation.

Define the 2 by 2 matrix, $A = \begin{pmatrix} 1 & \rho_{Y_bY_a} \\ \rho_{Y_bY_a} & 1 \end{pmatrix}$ and suppose that there are k covariates given by a vector, Z (i.e., the vector Z may include the variables age, baseline weight, etc.). Define the k by 2 matrix, $B = (\rho_{Y_bZ}, \rho_{Y_aZ})$ where ρ_{Y_bZ} represents a k by 1 vector of correlations between the outcome variable Y_b and each of the variables in Z , and ρ_{Y_aZ} is similarly defined. Finally define $C = (\rho_{ZZ})$ as the k by k correlation matrix of Z variables. Note that all correlations are estimable using observable data in the matrices B and C , but not the correlation in the matrix A . The positive definiteness requirement for a correlation matrix requires that the determinant $|A - B^TC^{-1}B| > 0$. Define the 2 by 2 matrix, $B^TC^{-1}B$, to

be of the form, $B^TC^{-1}B = \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}$. Then it can be shown that estimable tighter bounds for the non-estimable correlation are given by,

$$h_{12} - \sqrt{(1 - h_{11})(1 - h_{22})} < \rho_{Y_b Y_a} < h_{12} + \sqrt{(1 - h_{11})(1 - h_{22})},$$

and these tightened bounds lead to tighter bounds for σ_D .

Reference List

1. Segal NL, Allison DB. Twins and virtual twins: bases of relative body weight revisited. *Int J Obes Relat Metab Disord*. 2002; 26:437–441. [PubMed: 12075568]
2. Segal NL, Feng R, McGuire SA, Allison DB, Miller S. Genetic and environmental contributions to body mass index: comparative analysis of monozygotic twins, dizygotic twins and same-age unrelated siblings. *Int J Obes*. 2009; 33:37–41.
3. Bulik CM, Allison DB. The genetic epidemiology of thinness. *Obes Rev*. 2001; 2:107–115. [PubMed: 12119662]
4. Deram S, Villares SMF. Genetic variants influencing effectiveness of weight loss strategies. *Arq Bras Endocrinol Metab*. 2009; 53:129–138.
5. Qi L, Cho YA. Gene-environment interaction and obesity. *Nutr Rev*. 2008; 66:684–694. [PubMed: 19019037]
6. Loos RJ. Genetic determinants of common obesity and their value in prediction. *Best Pract Res Clin Endocrinol Metab*. 2012; 26:211–226. [PubMed: 22498250]
7. Sjostrom L, Peltonen M, Jacobson P, Sjostrom CD, Karason K, Wedel H, Ahlin S, Anveden A, Bengtsson C, Bergmark G, Bouchard C, Carlsson B, Dahlgren S, Karlsson J, Lindroos AK, Lonroth H, Narbro K, Naslund I, Olbers T, Svensson PA, Carlsson LM. Bariatric surgery and long-term cardiovascular events. *JAMA*. 2012; 307:56–65. [PubMed: 22215166]
8. Colquitt JL, Picot J, Loveman E, Clegg AJ. Surgery for obesity. *Cochrane Database Syst Rev*. 2009:CD003641. [PubMed: 19370590]
9. Rudolph A, Hilbert A. Post-operative behavioural management in bariatric surgery: a systematic review and meta-analysis of randomized controlled trials. *Obes Rev*. 2013
10. Franz MJ, VanWormer JJ, Crain AL, Boucher JL, Histon T, Caplan W, Bowman JD, Pronk NP. Weight-loss outcomes: a systematic review and meta-analysis of weight-loss clinical trials with a minimum 1-year follow-up. *J Am Diet Assoc*. 2007; 107:1755–1767. [PubMed: 17904936]
11. Senn S. Individual Therapy: New Dawn or False Dawn? *Drug Info J*. 2001; 35:1479–1494.
12. Poulson RS, Gadbury GL, Allison DB. Treatment heterogeneity and individual qualitative interaction. *Am Stat*. 2012; 66:16–24. [PubMed: 23204562]
13. Gadbury, GL. Subject -treatment interaction. In: Chow, S-C., editor. *Encyclopedia of Biopharmaceutical Statistics*. Informa Healthcare; London: 2010. p. 1316-1321.
14. Wang X, Elston RC, Zhu X. The meaning of interaction. *Hum Hered*. 2010; 70:269–277. [PubMed: 21150212]
15. Rubin DB. Estimating causal effects for treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*. 1974; 66:688–701.
16. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986; 81:945–960.
17. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*. 1985; 41:361–372. [PubMed: 4027319]
18. Peto, R. Statistical aspects of cancer trials. In: Halnan, KE., editor. *The Treatment of Cancer*. Chapman & Hall; London: 1982. p. 867-871.
19. Gadbury GL, Iyer HK, Allison DB. Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat*. 2001; 11:313–333. [PubMed: 12018782]
20. Svetkey LP, Harsha DW, Vollmer WM, Stevens VJ, Obarzanek E, Elmer PJ, Lin PH, Champagne C, Simons-Morton DG, Aickin M, Proschan MA, Appel LJ. Premier: a clinical trial of comprehensive lifestyle modification for blood pressure control: rationale, design and baseline characteristics. *Ann Epidemiol*. 2003; 13:462–471. [PubMed: 12875806]

21. Tchernof A, Despres JP. Pathophysiology of human visceral obesity: an update. *Physiol Rev.* 2013; 93:359–404. [PubMed: 23303913]
22. Efron, B.; Tibshirani, R. *An Introduction to Bootstrap*. Chapman & Hall; New York: 1993.

Figure 1.

A and B. Waist circumference loss (cm) as a function of baseline waist circumference (cm) in treatment and control participants in [20] for women (Panel A) and men (Panel B). Data source: Raw data obtained from [20].

Table 1

Estimated mean treatment effect, estimated bounds for treatment heterogeneity, and estimated bounds for the proportion of individuals responding opposite the mean effects (no covariates used).--Standard errors were estimated using 2000 bootstrap samples. Data source: Raw data obtained from [20].

	Mean effect	SD.min	SD.max	P.min	P.max
Female estimate	1.4329	1.6414	15.4661	0.1913	0.4631
Female SE	0.8735	0.7602	0.7497	0.1639	0.0232
Male estimate	4.0757	2.6949	10.9497	0.0652	0.3549
Male SE	0.7962	0.8410	0.8422	0.0473	0.0245

SD.min – estimated lower bound for standard deviation, $\hat{\sigma}_D^{\min}$

SD.max – estimated upper bound for standard deviation, $\hat{\sigma}_D^{\max}$

P.min – estimated lower bound for proportion responding opposite mean effects, $\hat{P}_{-D}^{\min}$

P.max – estimated upper bound for proportion responding opposite mean effects, $\hat{P}_{-D}^{\max}$

Table 2

Estimated bounds for treatment heterogeneity using only baseline waist circumference as a covariate. Data source: Raw data obtained from [20].

	SD.min	SD.max
Female estimate	2.7132	14.5799
Female SE	0.7955	0.7188
Male estimate	2.6930	10.9024
Male SE	0.7801	0.8271

SD.min – estimated lower bound for standard deviation, $\hat{\sigma}_D^{\min}$

SD.max – estimated upper bound for standard deviation, $\hat{\sigma}_D^{\max}$

SE – Standard error of the estimate

Table 3

Maximum and minimum bounds by gender for the standard deviation (SD) using baseline weight, baseline waist circumference, baseline body mass index, age, and race as covariates. Data source: Raw data obtained from [20].

	SD.min	SD.max
Females	3.33	14.02
Males	3.32	10.45

SD.min – estimated lower bound for standard deviation, $\hat{\sigma}_D^{\min}$

SD.max – estimated upper bound for standard deviation, $\hat{\sigma}_D^{\max}$