Obesity

Common Scientific and Statistical Errors in Obesity Research

Brandon J. George[1], T. Mark Beasley[2], Andrew W. Brown[1,3], John Dawson[4], Rositsa Dimova[5], Jasmin Divers[6], TaShauna U. Goldsby[1,3], Moonseong Heo[7], Kathryn A. Kaiser[1,3], Scott W. Keith[8], Mimi Y. Kim[7], Peng Li[1,2], Tapan Mehta[3,9], J. Michael Oakes[10], Asheley Skinner[11], Elizabeth Stuart[12], and David B. Allison[1,2,3]

This review identifies 10 common errors and problems in the statistical analysis, design, interpretation, and reporting of obesity research and discuss how they can be avoided. The 10 topics are: 1) misinterpretation of statistical significance, 2) inappropriate testing against baseline values, 3) excessive and undisclosed multiple testing and "*P*-value hacking," 4) mishandling of clustering in cluster randomized trials, 5) misconceptions about nonparametric tests, 6) mishandling of missing data, 7) miscalculation of effect sizes, 8) ignoring regression to the mean, 9) ignoring confirmation bias, and 10) insufficient statistical reporting. It is hoped that discussion of these errors can improve the quality of obesity research by helping researchers to implement proper statistical practice and to know when to seek the help of a statistician.

Obesity (2016) **24**, 781–790. doi:10.1002/oby.21449

Introduction

Obesity studies cannot advance without good science. Unfortunately, many recent articles have raised valid questions about the quality of science underlying some obesity research (1-21,23). Many of the flaws identified in these critiques stem from errors in statistical design, analysis, interpretation, and reporting. This article describes 10 errors that appear repeatedly in the literature and discusses ways to identify and correct such errors. Our purpose is to educate and enable researchers and reviewers to recognize these errors and to avoid making them in their own studies.

Significance of Statistical Tests

Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published." (Lykken, 1968) (23)

Statistical hypothesis testing

Consider the case in which an investigator is interested in contrasting the effects on body mass index (BMI) of an experimental diet compared with a control diet. At the end of the experiment, the researcher may see that the participants given the experimental diet had on average a 2-unit decrease in BMI, while those on the control diet had on average only a 1-unit decrease. Inference about whether the experimental diet resulted in a greater reduction in BMI depends greatly on the variability of the decrease among the subjects. Some subjects will have lost more weight than others, while others may have gained weight during the study. A subjective "eyeball" test for a difference between groups lacks scientific rigor, so researchers typically rely on statistical methods to make quantitative conclusions about their data.

Statistical hypothesis tests provide a framework for deciding whether observed values differ from what would be expected by chance under the assumption that the two treatments have identical effects (i.e., the difference in mean BMI changes between treatments is zero). The null hypothesis (H_0) is therefore written as

$$H_0 : \mu_{\text{exp}} - \mu_{\text{cntl}} = 0 \qquad (1)$$

where μ_{exp} and μ_{cntl} are the mean population changes in the outcome of interest (here, BMI) for the experimental and control groups, respectively. The alternative hypothesis (H_A) is simply that the treatment means are different, which is written as

[1] Office of Energetics, University of Alabama at Birmingham, Birmingham, Alabama, USA. Correspondence: Brandon J. George (brgeorge@uab.edu) [2] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama, USA [3] Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, Alabama, USA [4] Department of Nutritional Sciences, Texas Tech University, Lubbock, Texas, USA [5] Department of Biostatistics, University at Buffalo, Buffalo, New York, USA [6] Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA [7] Department of Epidemiology & Population Health, Albert Einstein College of Medicine, Bronx, New York, USA [8] Division of Biostatistics, Department of Pharmacology and Experimental Therapeutics, Thomas Jefferson University, Philadelphia, Pennsylvania, USA [9] Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, Alabama, USA [10] Department of Epidemiology & Community Health, University of Minnesota, Minneapolis, Minnesota, USA [11] Department of Health Policy and Management, University of North Carolina, Chapel Hill, North Carolina, USA [12] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.

See Commentary, pg. 767.

Funding agencies: Funded in part by NIH grants P30DK056336, R25HL124208, R25DK099080, R01MH099010, R01MD009055, R25GM116167, and T32DK062710. The opinions expressed are those of the authors and not necessarily the NIH or any other organization.
Disclosure: The authors declared no conflict of interest.
Received: 29 October 2015; **Accepted:** 7 December 2015; **Published online 30 March 2016. doi:10.1002/oby.21449**

TABLE 1 Possible outcomes of a frequentist, 2-sided hypothesis testing exercise

Unobserved reality in the population ("Truth")	Result of test based on observed data	
	Do not reject H_0	Reject H_0
The 2 treatments are equivalent H_0 is true	Correct	Type I error
The 2 treatments are different H_A is true	Type II error	Correct

$$H_A : \mu_{exp} - \mu_{cntl} \neq 0 \quad (2)$$

There are four possible outcomes of a hypothesis test as shown in Table 1. In two cases, the researcher gets it right: not rejecting H_0 when the treatments are equivalent and rejecting H_0 when the treatments are different. But the researcher can also get it wrong in two ways: rejecting H_0 when the treatments are actually equal, which is known as a Type I error or false positive, and failing to reject H_0 when the treatments are actually different, which is known as a Type II error or false negative.

Defining null-hypothesis significance testing

Significance testing is often discussed in terms of P-values. The P-value is the probability (ranging from 0 to 1) of observing a given result (or something more extreme) under the assumption that H_0 is true. In the above example, it would be the probability of seeing the difference in BMI means between the groups by chance, if in fact the diets have no difference in average effect (24). The reader can then decide whether the results are statistically significant on the basis of a prespecified significance level (α) of their choosing, in which the null hypothesis is rejected if the P-value is less than α. This approach is often referred to as null-hypothesis significance testing (NHST) and its limitations and statistical problems have been described in great detail elsewhere (23,25,26). Despite the shortcomings of NHST, the use of significance testing is engrained in biomedical research and will likely remain in use for the foreseeable future. The rest of this section provides guidance for reducing the likelihood of reaching the wrong conclusion when using NHST.

Null-hypothesis significance testing in practice

One of the key criticisms leveled against NHST is that it does not answer the right question. The investigator really wants to know the probability that H_0 is true given the collected data (e.g., the two diets have the same effect on BMI), but NHST only provides the probability of seeing the collected data given that H_0 is true in the population. NHST is therefore rooted in proof by negation, i.e. these data are unlikely given that there is no actual difference. Proofs by negation can be valid but are unreliable when the premise is based on probabilities. This limitation means that NHST cannot prove the strict "if A then B" relationship that researchers are often interested in.

Authors sometimes mistakenly assert that a large P-value (e.g., near 1) provides evidence that H_0 is true (e.g., that two groups are equivalent), when in fact it only suggests that there is insufficient information to reject the null (27). Compare this to a court finding someone "not guilty" instead of "innocent": the person is first assumed innocent, and there is not enough evidence to find them guilty. If a researcher wishes to show a lack of a difference between groups, one can perform an *equivalency study*. However, one cannot test equivalency in a *post hoc* manner (28) and due to the complexities of the design collaboration with a statistician is essential. Also note that the standard NHST approach considers a two-tailed test, where the alternative hypothesis allows for the mean difference to be either positive or negative. One-tailed tests (that test for a difference in just one direction) are inappropriate in most circumstances.

The role of effect size in hypothesis testing

Errors in interpreting significance go beyond the classic issues of Type I and Type II errors, including mistakes in interpreting the experimental results that can occur if the investigator focuses solely on the results of NHST while ignoring the magnitude of the association, commonly referred to as the *effect size*.

A common case of this problem is touting the importance of a statistically significant difference despite the estimated effect size being scientifically insignificant. For example, one study concluded that snacking "independently contributes to hepatic steatosis and obesity" (29). However, the results showed an increase of only 0.8% in intrahepatic lipid content in the experimental group compared with the control, with both groups remaining well within clinically acceptable liver lipid concentrations.

Ignoring implausibly large magnitudes can also occur. Schoenfeld and Ioannidis documented the effect sizes reported for associations between reported food intakes and cancer and noted "implausibly large effects, even though evidence is weak" (30). "[Taken] literally-if we increase or decrease (as appropriate) intake of any of several nutrients by 2 servings/day, cancer will almost disappear worldwide" (7). The feasibility of results should be thoroughly considered during the writing and review of the studies.

Some researchers focus primarily on the P-value and whether it is under the ubiquitous significance level of 0.05. To provide greater information and interpretability, additional statistics including confidence intervals and measures of observed effect sizes must always be reported (31). For example, statements of the form: "Treatment A led to a 0.5-standard deviation reduction in Y" can help readers judge the likelihood of a change of this magnitude in the population of interest and whether the resulting change in the outcome is relevant in practice.

A larger *n* may not lead to "better" results

A common mistake in thinking about statistical significance and sample size is to assume that results only get better (i.e., smaller P-values) by increasing the sample size. A common refrain when a test is "close to significant" (e.g., $P = 0.051$) is to suggest that a larger sample size would have made it significant. Here, we discuss why an increase in sample size may not lead to statistically significant results.

Suppose that an investigator is testing the null hypothesis $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, where μ_0 is a constant. This test can be based on

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \qquad (3)$$

where $\bar{X}$ and S represent the observed mean and standard deviation of the variable of interest whose mean is being tested using a sample of size n. An increase in sample size may not lead to an increase in T (and thus a decrease in the P-value) if the absolute difference between $\bar{X}$ and μ_0 decreases (resulting in a small numerator) or the sample standard deviation S increases (larger denominator) in the new sample. This behavior can be even more unpredictable when the original sample is not representative of subsequent samples.

Difference in Nominal Significance is not a Significant Difference

Randomized controlled trials (RCTs) are comparative studies in which subjects are randomly assigned to receive either the intervention(s) or the control (placebo or current standard intervention) with the hypothesis that the novel intervention will have an effect on a specific outcome (e.g., body mass, percent fat mass). The randomized group allocation is intended to produce comparable groups, such that measured and unknown subject characteristics and variables at the time of randomization, on average, are balanced between the groups. Typically, the study outcome is measured at baseline and again at the end of the trial after a prespecified follow-up period.

A frequently encountered error in the obesity literature involving parallel group RCTs with pre- and post-intervention data is the use of within-group paired tests as opposed to between-group tests. Here, researchers base their inference on the difference in significance of the outcome between the pre- and post-intervention measurements rather than the significance of the difference between groups. For example, Cassani et al. recently described an RCT of the effect of flaxseed consumption on body weight and inflammatory markers (32). The study incorrectly concluded that because the inflammatory markers C-reactive protein and TNF-α decreased significantly in the flaxseed group only, adding flaxseed to a weight-loss diet could reduce these markers. The appropriate analysis revealed no significant difference in the final outcomes between the two groups (33). This type of erroneous inference was described by Boutron et al. as one of the often encountered “strategies for spin” in reports that typically focus only on the statistically significant results (34). This within-group approach is invalid and can have a false-positive rate for detecting a difference of up to 50% for two treatment groups (and potentially higher for more than two groups) (35). This is because the difference in nominal significance and its true false-positive rate are actually functions of the power of the paired tests to detect a pre–post difference within each group rather than any difference in the effect of the intervention on the outcome’s change over time. In less quantitative terms, if each arm has 50% power to detect a pre–post difference, then this analysis approach is equivalent to flipping two coins and declaring a difference if they did not land on the same side.

In terms of practical interpretation, this analysis strategy does not make sense. Consider two arms in a weight-loss trial: one with a 95% confidence interval of (0.1 kg, 1.9 kg) for a 6-week weight loss and the other with interval (−0.1 kg, 2.1 kg). The first arm experienced a nominally significant weight loss over the 6-week period while the second did not, but both arms report an average weight loss of 1 kg over the 6-week period and the intervals are not meaningfully different.

There are several ways to assess the treatment effect in parallel group RCTs that involve pre- and post-intervention evaluation of the outcome (36). The first, referred to as *endpoint analysis* or *change score analysis*, involves calculating the change from baseline to follow-up for each subject and running a two-sample *t*-test (two treatment arms) or analysis of variance (ANOVA; three or more arms) with the observed change as the measured outcome. The second method, described as a *baseline-adjusted analysis of covariance* (ANCOVA), analyzes the data in a linear model with the subjects’ follow-up values as the outcome and the treatment and observed baseline values as the independent variables. The second method is readily available in statistical software, is straightforward to run, and typically has more power than endpoint analysis (37-39). Although more complicated methods of analysis exist for this type of data (36,40-42), the common theme among all proper methods for testing a treatment effect over time is that the actual difference in the change over time is tested between groups.

In summary, a researcher should *never* use the nominal significance of a pre–post difference within a group to make inferences about differences between groups.

Multiple Testing and *P*-Value Hacking

Data from RCTs and observational studies are often analyzed using the NHST framework as part of the confirmatory analysis. *Confirmatory analysis* implies inferential analysis in which the variables, model, and NHST to be conducted are specified before looking at the data. Although in theory the Type I error rate of a NHST should be governed by the costs of falsely rejecting a true hypothesis, the *de facto* probability level in the literature is 5%.

Multiple testing refers to testing more than one hypothesis at a time (43). One of the principles overlooked is that when many hypotheses are tested the probability of getting at least one false-positive increases. That is, the multiple tests lead to an inflated Type I error rate unless correction procedures are applied. There are different error rates, such as the false-discovery rate, error rate per hypothesis, error rate per family, and family-wise error rate, and the choice of error rate should depend on the experimental situation. For example, in high-dimensional genomic studies where the cost of a Type I error is not as large as in an intervention testing a drug or policy, some authors have recommended using the false-discovery rate (44). New methods are also being proposed to identify the right level of Type I error for a given study after accounting for the cost of a Type I error (45,46).

In this section, we focus our attention largely on the family-wise error rate, defined as the probability of at least 1 Type I error in the family. The practice of testing several hypotheses while controlling the family-wise error rate raises the question, What is a “family” of hypothesis tests? A “family” of hypotheses can be defined in at least two logical ways: either in terms of testing several different outcome measures for a given intervention or risk factor or in terms of comparing several interventions for a single outcome measure. For example, if a diet and lifestyle weight loss study had primary outcomes of weight, visceral adiposity (via MRI), and glycemic control (via HbA1c), then the three can together be considered a family of hypotheses. Furthermore, within a given experiment, the investigators may be testing the efficacy of multiple interventions (e.g., different diets). In this scenario, the hypothesis tests used to estimate the efficacy of multiple intervention arms compared to the control group constitutes a family of hypotheses. Regardless of the definition used, the family of hypotheses needs to be predefined and independent of the analysis results. For RCTs following the CONSORT guidelines, the definition of a family for the primary analysis should be noted in the trial registration prior to the study (47,48). For secondary data analyses of existing RCT or observational data, for which guidelines (e.g., STROBE) to register analyses have not yet received widespread acceptance, this has come to be known as *P-value hacking* (49,50).

Back in 1993, Mills noted, "If you torture your data long enough, they will tell you whatever you want to hear" (51). *P*-value hacking, in which investigators run different forms of analysis on a data set until they find results that suit them, is one such practice of torture. More commonly, the preference is to reanalyze data until a "statistically significant" result is discovered under the assumption that validly conducted NHST with nonsignificant results will not lead to publication (52,53). A classic example of *P*-value hacking is the practice of identifying subsets of data that lead to significant findings. The findings for these subsets are then reported as if this was the central question of the study.

Another way *P*-value hacking is introduced is through model selection procedures such as stepwise regression. An important step in conducting NHST is choosing the covariates to be included in a model, which ideally should be prespecified and based on some rationale derived from existing knowledge. However, this aspect is often relegated to a model selection process such as stepwise regression, a sequential variable selection procedure for building a model in which variables are added or removed based on stopping rules that may be based on the significance of the fitted model (54). Implicit in stepwise regression is a process of multiple hypothesis testing, which is often ignored. Thus, the resulting model is not only overfitted to the data but also has an inflated Type I error rate. Such model selection techniques implicitly promote *P*-value hacking and should only be applied as exploratory techniques to identify potential predictors from a test data set to be validated in a different experimental data set (55).

There is debate in the field regarding whether multiple testing correction should be standard or if it should be left to the reader. Although we cannot make a definitive statement on whether it should be done, we do feel that failing to disclose multiple testing, particularly iterative analysis strategies based on significance levels like *P*-hacking, is an error. We feel that authors must be clear about how many tests were run and how they came to their conclusions so that readers can make informed interpretations of findings.

Cluster Randomized Trials

A cluster randomized trial (CRT) is an experimental design in which specific social groups called clusters are randomized to interventions (56). Commonly examined clusters include schools, clinics, and neighborhoods; common interventions include changes to food environments and policy changes. CRTs are not ecological designs, although they share some characteristics. In fact, a distinguishing characteristic of CRTs is that while randomization happens at the cluster level, the outcome variables (such as BMI) are measured at the individual level. CRTs are not multisite RCTs, in which persons *within* a cluster are randomized to treatment conditions; in CRTs all members of a given cluster are treated or not. In CRTs, subjects are nested within clusters, and clusters are nested within experimental conditions (57).

Clustering arises because persons within a given cluster are typically more alike than persons between clusters. Consequently, there is less independent information within a cluster than the total number of subjects, meaning the effective sample size is smaller than the actual sample size (57). Consequently, specific methodologies are required to draw credible conclusions from CRTs.

Due to clustering, the total variance of the outcome variable consists of between-cluster variance and within-cluster variance. The intraclass correlation (ICC), typically denoted by ρ, is the ratio of the between-cluster variance to the total variance of an outcome variable and is viewed as a measure of the strength of clustering (58). The design effect (DEF) is the ratio of the variance estimate of an outcome variable taking the clustering into account to the variance estimate of the outcome variable ignoring the clustering and can be calculated as

$$\mathrm{DEF} = 1 + (m-1)\rho \qquad (4)$$

for clusters of equal size m. It must be noted that the number of subjects is not necessarily the same in all clusters in CRTs; unequal cluster sizes may lead to larger design effects and are therefore less efficient than equal cluster sizes provided the same total number of subjects, though this can rarely be controlled in a study (56).

Ignoring nesting and clustering effects in CRTs will cause inflated Type I and Type II errors by underestimating the variance of intervention effects and overestimating the degrees of freedom (df) in the hypothesis testing (59). Some researchers incorrectly state that a small ICC will not inflate variance estimates too much and therefore claim it is not necessary to take clustering into account. However, the design effect can still be large given a small ICC if the cluster size is large, which is the common situation in CRTs. For example, in a school-based obesity intervention trial with an average of 100 students per school and an ICC of 0.01, the design effect will be 1.99, which suggests that the "true" variance will be twice as large as the variance estimate when clustering is not taken into account.

Similarly, the df available for statistical inference is limited by the number of clusters in the study. For a hypothetical two-armed CRT with $K = 10$ clusters in each arm and $m = 100$ subjects in each cluster, there will be $N = 2Km = 2000$ total subjects. In this case, the available df to test the intervention effect is only $2(K - 1) = 18$ rather than $N - 2 = 1998$. Although the N can be very large, the fact that there are typically only a small number of clusters in CRTs makes the df for hypothesis testing of intervention effects very limited. Consequently, a test statistic using appropriate denominator df of $2K - 2$, such as a t test or an F test, should be used. Other tests, such as χ^2 or z tests based on large sample approximations, can lead to false-positive results as they assume infinite df. Many software packages do not provide the correct df automatically; special programming is required to obtain the proper df in hypothesis testing, even in models in which the ICC is correctly considered for the variance estimation.

An extreme case is the "one-cluster-per-condition," i.e., $K = 1$, in which no valid statistical inference on intervention can be obtained regardless of cluster size because of the zero df [$\text{df} = 2(1 - 1)$]. In other words, the intervention effect cannot be differentiated from the cluster effect. Therefore, the one-cluster-per-condition design should be avoided unless one is collecting pilot data to estimate the ICC for the power calculation of a subsequent CRT.

In summary, CRTs require special analytical methods to account for the within-cluster correlation and the df being limited by the number of clusters rather than the number of subjects, even when the correlation is hypothesized to be small. Understanding of these points is critical in designing and analyzing CRTs for valid statistical inferences, and consulting from experienced statisticians is highly recommended.

Misconceptions of Nonparametric Tests

Some of the most common misconceptions in statistical practice are that nonparametric tests are "distribution free," "assumption free," or less powerful than their parametric counterparts. We will use the Kruskal–Wallis (KW) test (i.e., Mann–Whitney *U*, Wilcoxon rank sum) as an example of a nonparametric counterpart to the parametric independent samples *t*-test or one-way ANOVA; however, these misnomers also occur for many other nonparametric procedures.

The rank-based KW test is a variation of performing an ANOVA (a parametric test) on ranks (e.g., the lowest value is 1, the second lowest is 2, and so on). For smaller sample sizes, an exact *P*-value can be computed based on the permutation of the ranks without assuming a distribution for the outcome or residuals. As sample size increases, the permutation distribution becomes more difficult to compute since the number of permutations for a one-way design with *J* groups is

$$\frac{N!}{n_1!n_2!\cdots n_j!\cdots n_J!} \tag{5}$$

where $\sum_{j=1}^{J} n_j{=}N$ is the total sample size. For two groups with sample sizes of $n_1 = 7$ and $n_2 = 6$, the number of permutations is $13!/(7!6!)\ = 1{,}716$, which is large but still computationally feasible. For a study with 30 subjects in 3 groups ($n_1 = n_2 = n_3 = 10$), there are approximately 5.55 trillion possible permutations. Thus, it does not take a large or complicated study to make the calculation of exact *P*-values unreasonable. Fortunately, as the sample size increases the permutation distribution underlying the test approximates a scaled chi-square distribution with $J\ -1$ df (60). Therefore, for designs with a prohibitively large number of permutations, researchers can use the KW approximate chi-square test but doing so requires the assumption that the test statistic has a parametric distribution.

The assumptions of the parametric ANOVA model involve normal independent distribution of errors with a mean of zero and a constant variance for all observations (homoscedasticity), denoted as NID$(0,\sigma^2)$. For the KW test, if the observations are not independent [i.e. repeated measures (61,62) or clusters (63)] the permutation distribution is invalid and other rank-based procedures must be used. In the strictest sense, rank-based statistics test a null hypothesis of two groups having the same distribution (64-66), although the rank-based tests are especially sensitive to differences in location (i.e., one distribution is shifted up or down) (67). If one assumes that the errors are independent and identically distributed [IID$(0,\sigma^2)$] but not necessarily normal, then rank-based procedures test differences in location (64). Zimmerman (1996) showed that heteroscedasticity in the original data will be inherited by a rank transformation (68), therefore unequal sample sizes and unequal variance will also affect the Type I error rates of rank-based tests (69).

In terms of statistical power, many statistics texts note that the asymptotic relative efficiency (a measure of power) of a rank-based test relative to the parametric *t*-test is $3/\pi = 0.955$ under the parametric NID$(0,\sigma^2)$ assumptions. But, for many asymmetric, homoscedastic error distributions [IID$(0,\sigma^2)$], rank-based tests have more power than their parametric counterparts (70). Furthermore, if the null hypothesis of the parametric *t*-test is true (i.e., the means are equal) but the groups differ in their distributional shapes, the rank-based test will have statistical power to detect a difference between the groups. To illustrate, suppose a control group has a positively skewed distribution with a mean BMI of 26 kg/m^2 and a standard deviation of 3. Suppose a treatment group with a negatively skewed distribution, mean of 26 kg/m^2, and standard deviation of 3. These two groups will differ in their 1) distributional shape, 2) medians (and other quantiles), and 3) number of high scores (which will be larger in the treatment group). In this situation, the null hypothesis for the parametric *t*-test is true (i.e., the means are identical) but the distributions are not identical and a rank-based test will detect these types of distributional differences. If researchers are comfortable with not focusing on a single parameter (i.e., mean difference) and rejecting a different null hypothesis of "stochastic heterogeneity" (i.e., the distributions are not equal), then non-normality and heteroscedasticity are not necessarily "nuisances" or "violations" but rather a part of the results (66).

In summary, nonparametric tests are not necessarily "distribution free" because large-sample approximate tests are assumed to follow known distributions. Nonparametric procedures may "relax" some of the parametric assumptions; however, they are not "assumption free" because the independence assumption is crucial to the permutation distribution underlying the test statistics. They are also not necessarily less powerful than their parametric counterparts and in some cases are more powerful.

Handling of Missing Data

Missing data are ubiquitous in research: people drop out of randomized trials or don't respond to individual survey questions. Incorrectly handling missing data can yield incorrect research results. We present solutions to missing data that can be used in practice.

A common (but inappropriate) approach to handling missing values is to simply ignore them, in the sense of dropping from the analysis individuals who have missing values for the variables of interest.

This is called *complete case* or *listwise deletion* and is often the default in statistical software packages. Unfortunately, complete case analyses have reduced statistical power and may yield incorrect answers if missing values systematically differ from observed values. For example, in a randomized trial of an obesity reduction program, individuals who had high BMIs at baseline or who were not successful at losing weight may be more likely to drop out of the study. Analyzing only those individuals with observed outcomes may not yield results applicable to the full study population.

A second common, but generally inappropriate, strategy is *single imputation*, whereby researchers "fill in" (or "impute") the missing values once. Types of single imputation include mean imputation, regression-based imputation, hot-deck imputation, last observation carried forward, and a missing data indicator approach. None of these is an appropriate strategy (71,72), and they are listed only so they can be recognized and avoided. Single imputation approaches yield incorrect results as they will overstate the precision of the study results: standard errors will be smaller than they should be and *P*-values will be more significant than they should be. This is because single imputation does not account for the uncertainty in the imputations: the analysis is done as if the imputed values are the true, actual values, when in fact they are not.

A better, and very flexible, approach for handling missing data is *multiple imputation*, which essentially repeats the regression-based or hot-deck single imputation approaches (which use observed values and known covariates to predict the unobserved values) multiple times. Multiple imputation creates multiple "complete" data sets. Analyses are then conducted separately within each data set and the final results are obtained by combining the data set-specific estimates using established formulas (73). The key distinction from single imputation is that the variance accounts for the variability within each data set as well as across imputations. The result is accurate standard errors that account for the uncertainty in the imputations. Software for creating multiple imputations and analyzing multiply imputed data is now readily available in SAS, Stata, and R, among other packages (see, e.g., http://www.stefvanbuuren.nl/mi/Software.html). Maximum likelihood and weighting approaches can also be appropriate methods for handling missing data in some analyses and data settings but are less versatile than is multiple imputation. See White et al. (2011) or Carpenter and Kenward (2013) for a summary of practical considerations in conducting analyses using multiple imputation, including implications for randomized trials (74,75).

Of course, avoiding missing data in the first place is the best strategy (71). Best practices for research should include 1) limiting missing data, 2) documenting the extent of missing data and exploring the reasons for missing data, and 3) prespecifying, and then using, an appropriate method for handling missing values, such as multiple imputation.

Calculation of Effect Sizes for Meta-Analyses

Meta-analyses place literature reviews on objective quantitative ground by calculating a formal quantitative measure of the magnitude of the effect or association under study, called the *effect size*. We use the term effect size to broadly denote any quantitative measure of the estimand under study in a meta-analysis and its use does not necessarily imply a cause and effect relationship. There are many measures that could be considered an effect size (e.g. Pearson correlation, Cohen's *d*, Phi coefficient) but no single one is best in all situations or is always appropriate. Almost all effect sizes can be transformed via some calculations to another with a one-to-one relationship (76-78), though one should note that different measures may not be directly comparable (particularly if the variables under study have different forms).

Common effect size errors

Although the pooling of effect sizes is relatively easy with meta-analysis software, that ease may belie the complexity of the underlying decisions and procedures involved in properly calculating the effect sizes which could lead to errors.

Perhaps the most thorough analysis of common errors in the calculation of effect sizes in meta-analysis was published by Gøtzsche et al. (79). Gøtzsche et al. studied published meta-analyses that used standardized mean differences (SMDs, sometimes referred to as Cohen's *d*) and spot-checked the calculations against the original papers. Of the 27 meta-analyses considered, 10 were found to have discrepancies in SMD calculation, 17 contained errors, and 3 had the overall conclusion change or be refuted to the point of retraction.

In our own work of reading published meta-analyses, we have also observed that errors in the calculation of effect sizes are common (80) and often fall into two major categories. The first involves a faulty imputation procedure. The second involves miscalculation of or incorrect choice of a variance.

With respect to imputation procedures, if an original article does not report the data to calculate an effect size exactly (e.g., reporting a result as "non-significant"), there are ways to try to back-calculate (81) or as a last effort impute the missing data (82). What should not be done, however, is to set the effect size of a non-significant effect to zero or choose an arbitrary *P*-value greater than 0.05 (83,84). These errors are reminiscent of missing data issues discussed in the previous section *Handling of Missing Data*.

With respect to variance, the challenge seems to be in understanding which variance is desired in a particular situation; this affects not only the calculation of SMDs but also the variance of the meta-analyzed effect size itself and weighting factors used. Most often, the variance desired is the within-group among-subject variance in the outcome measure. This value can be used to standardize different measurement scales to construct SMDs (85). However, a meta-analyst can run into trouble if different types of variance are used to scale the SMDs, particularly when the studies in the meta-analysis have a pre–post design. In this case, the several different variances (e.g., post only, pre and post pooled, post minus pre) one could choose to use as a denominator for an SMD (86) have different implications. The appropriate choice of variance is context-dependent, so it falls to researchers to consider which has the most scientific merit for their topic and report their choice and rationale.

One must also take care in determining whether covariates were included in the analysis reported by a study. Although in RCTs covariates should not affect the raw size of an effect (e.g. mean difference) (87), their inclusion may reduce the residual variance of the overall model (39). Combining an effect size calculated from this reduced residual variance with effect sizes calculated from studies that did not include the same covariates is comparing apples and oranges. Such errors are discussed elsewhere (83).

The final situation we will consider in which choice of the variance seems to confuse many investigators involves CRTs (3,88,89). The key point of confusion seems to be whether to use the within-condition among-cluster variance or the within-population among-subject variance for the meta-analysis. Although the former may be appropriate for NHST (3), if one is including CRTs and ordinary RCTs in a single meta-analysis the latter is appropriate.

Recommendations

First, investigators should specify how the effect sizes and corresponding standard errors were calculated and do so with greater completeness and precision than is common practice today. Second, we suggest that doing meta-analysis well requires collaboration with someone with advanced training in meta-analytic calculations.

Ignoring Regression to the Mean

Regression to the mean (RTM) is a statistical phenomenon that occurs when repeated measurements are made on the same subject or same unit of observation. When observed values have random error (i.e., nonsystematic variation around the true mean), subsequent observations tend to regress to the population mean. Francis Galton first recognized this phenomenon over a century ago and described it using the heights of parents and children (90). In this classic description, when parents were taller than the population average, their children tended to be shorter than the parents (regressed down toward the mean), and when parents were shorter than average the children tended to be taller than the parents (regressed up toward the mean).

RTM can be worsened when categorizing subjects on the basis of baseline measurements for two reasons (91). For example, blood pressure is often categorized as normal, pre-hypertension, and hypertension (92). First, when each category is defined on the basis of its distance from the population mean, examining only subjects at the extremes (e.g., those with hypertension) will always result in greater RTM than examining subjects nearer to the mean. Investigators may not find an overall group effect, but will when they sample the population asymmetrically (93). Second, if one measure of blood pressure is either much higher or much lower than the mean, a second measure will likely be closer to the mean. An average of multiple measures can be used to reduce RTM due to random variation. However, RTM also represents a change in the true value of a variable over time, not only random variation. An individual whose true blood pressure is high will tend to have a lower blood pressure at subsequent time points. Whenever two variables are not perfectly correlated, true values will always regress to the mean regardless of measurement error, the order of measurement, or whether the two variables are repeated measures of the same construct.

One of the most common errors associated with RTM, particularly in obesity literature, is concluding that an intervention is effective when the study design does not permit such a conclusion. School-based interventions appear particularly susceptible, because RCT designs are less common. These interventions often ignore the effect of RTM, reporting reductions in BMI *z*-score (94,95) and prevalence of obesity (96) compared only to baseline. Community-based interventions make similar claims based on comparisons to baseline, reporting success in reducing weight and blood pressure (97) even when lacking a control group (98).

Another interpretation error that can be caused by RTM is an assumption of differential treatment effects based on baseline values of the outcome variable. For instance, studies may use greater declines in BMI among participants with higher baseline BMI than those with lower baseline BMI as evidence for differential treatment efficacy by baseline BMI (99,100). Differential weight changes as a function of baseline BMI would be expected solely from RTM. The only way to determine differential treatment based on baseline values is through the use of a control group and testing for an interaction between baseline value and treatment.

The clearest way to avoid RTM leading to unsubstantiated inferences about efficacy is through the use of an appropriate control group, ideally by random assignment of the study subjects. RTM then becomes an untenable explanation for any difference in outcome between the two groups (101). When true RCTs are not feasible, reasonable alternatives should be implemented. For example, if equity concerns are raised in the randomization of a school-based intervention, alternatives such as crossover trials should be considered. Nonrandomized comparator groups, imperfect but still useful, can be identified through alternative designs, such as use of a contemporaneously measured unexposed cohort with similar characteristics. Finally, quasi-experimental designs provide stronger evidence than do uncontrolled interventions in which investigators simply look at change from baseline and a group of treated cases (102). Ignoring the potential effects of RTM can lead to unsubstantiated inferences about the effects of treatments that can lead to wasted time, money, and other societal resources and distract from alternative interventions that may be more valuable.

Ignoring Conformation Bias

Sackett, regarded as the father of evidence-based medicine (103), listed an expansive catalog of sources of bias that can occur at each of the seven general stages of research, from “reading-up on the field” to “publishing the results” (104), and may lead to statistical and other inferential errors. However, one important source of bias is missing from Sackett’s list: confirmation bias. This bias is the tendency for researchers to evaluate analytical results less critically when the results are consistent with their prior beliefs of the study outcome or the hypotheses they are aiming to prove. Well-formulated hypotheses based on past research experiences and findings are crucial for advancing knowledge. At the same time, a very strong scientific rationale or premise for conducting a new study may lead to overconfidence in one’s current findings if they are in the expected direction, leading investigators to check them less thoroughly than results that are unfavorable or counterintuitive to their hypothesis. On the other hand, if results are not in the expected direction, confirmation bias can place improper influence on the conduct of data analysis.

Confirmation bias can be manifested in many ways, including failure to identify data entry, coding, programming, and other errors, as well as overlooking the potential for confounding in the experimental design or analysis. With regard to the latter, observational studies may be more vulnerable to the effects of confirmation bias than RCTs. For example, an observed association between incidence of lung cancer and weight could be biased if smoking status was not well controlled. Similarly, in genetic studies, ignoring population stratification could lead to the wrong conclusion regarding the effect of a locus on risk of disease.

Randomized clinical trials are not entirely immune to the effects of confirmation bias. For example, suppose that a randomized trial is conducted to evaluate the impact on weight loss of a lifestyle intervention and that intake of weight-loss pills was allowed for ethical reasons in both the intervention and control arms. If the intervention group had a higher rate of weight-loss pill intake than in the control group, the effect of the intervention would likely be overestimated. Insufficient control for use of the weight-loss pill due to confirmation bias may lead to erroneous claims about the benefits of the intervention.

How can we prevent or minimize such confirmation biases in research? The following suggestions, although perhaps onerous and far from exhaustive, may help: 1) At the design stage, statisticians should communicate closely with investigators to thoroughly understand the research setting; identify potential sources of error, confounding, and bias; and determine the most pertinent statistical methods and analytic strategies for the study. Also, sensitivity analyses should be an important component of any statistical plan to ensure that study findings are robust to different analytic approaches. 2) In the analysis, the statistician should adhere as closely as possible to the analytical plans that were determined *a priori*, and investigators should ideally not be involved in the data analysis. To further ensure that data are properly handled and analyzed, large observational studies should have an independent data monitoring committee the way most clinical trials do. Most importantly, data analysis should not be guided or influenced by the indications of results. 3) Whenever possible, parallel data analyses by an independent statistician should be conducted to confirm the results of the study statistician. 4) Each aspect of data collection, storage, cleaning, analysis, and output should be logged and archived for review to enhance the reproducibility of the analytic results. These steps are also consistent with the recent recommendations by the National Institutes of Health to enhance scientific reproducibility and transparency through rigorous experimental design, appropriate analytic approaches, and other sound statistical practices (105).

As Popper stated, “Those among us who are unwilling to expose their ideas to the hazard of refutation do not take part in the scientific game” (106). All of us should be open to such refutability of study findings, perhaps for the very reason that no study findings may be free from any sources of bias including confirmation bias.

Errors in Reporting

Here we briefly outline types of common reporting errors: 1) insufficient detail on methods and results, 2) unclear statement of primary and secondary outcome variables, and 3) the use of causal language without appropriate data to support such conclusions.

An error of insufficient detail occurs when pertinent information has not been included in the manuscript, hampering a reader’s ability to check the validity of the analysis and a researcher’s ability to include the work in a meta-analysis. The most common of these errors involve insufficient precision in reported values or P-values, such as reporting an odds ratio of 1.3 (95% confidence interval: 1.0, 1.5) or reporting a P-value as NS (not significant). Although it is not useful to give overly precise values, inherently unstable estimates such as odds ratios would benefit from two to three decimal places at a minimum. Fortunately, this is an easy problem for an author to correct.

Additionally, omission of details about the exact modeling approach is common and easily avoidable. Statistical analysis sections of papers should not only report what software was used to analyze the data but also the exact routines and options such that the analysis could be reproduced by a third party. The needed level of detail includes the exact variables used and how the variables were used in the model. Most desirable would be the inclusion of syntax in an online supplement or appendix. Simply stating that “the general linear model was used in SAS” is inadequate.

Another reporting error is the lack of clarity on outcomes of interest. In this era of “big data,” researchers can increasingly assess many different potential outcomes without clear hypotheses *a priori*. The choice of the outcomes to highlight is one of the many so-called “researcher degrees of freedom,” coined by Uri Simonsohn (107) and popularized by Gelman (108,109), that refers to the many choices that researchers make during the design, execution, and analysis of an experiment that may impact the results.

The last form of reporting errors we will discuss is the use of causal language in nonrandomized studies. Particularly in obesity research, scientists sometimes attempt to explain the mechanisms behind observed associations. However, nonrandomized studies can, at best, only provide information about correlations among the variables, not causality. While it may be tempting to imply that the outcome is logically or chronologically subsequent to the variables being treated as its predictors, support for such an assumption cannot come from the data values themselves. While we may assume that gene expression may be driving a phenotype and not vice versa, we may not claim that consumption of a particular item causes obesity merely because the two are associated in cohort studies. Rather, one of these cases might be in play:

- Reverse causality: Instead of X causing Y, Y may cause X.
- Third-variable situations (nonmediators): Some unobserved variable M may be causing X and may be causing Y. Without conditioning on M, there will be an association between X and Y.
- Third-variable situations (mediators): X causes M and M causes Y, so M is mediating the effect of X on Y. Distinguishing effects of third-variables requires support from additional data that can elucidate the causal and temporal pathways between the variables (110).

Thus, the use of causal language such as ‘the effect of,” “causes,” or “influences” is not appropriate when discussing nonrandomized studies. Softening phrasing, such as “may cause,” does not ameliorate this concern. Even a phrase such as “is linked to,” which properly denotes association, has causal connotations and should be avoided. Stronger statements of the limitations of the data and the conclusions that can be drawn are needed, even when biological plausibility exists.

Discussion

Considering frequent appearance of these errors in the literature, it is clear that obesity researchers need more rigorous statistical support and training. This could come from course work during graduate or postdoctoral training or from workshops or short courses. Utilization of published guidelines such as CONSORT (48) or PRISMA (111) may also be useful for producing valid research. Furthermore, we speculate that including a statistician in both the research and reporting stages of the scientific process may produce higher quality, more valid, and more reproducible results. We hope that tutorials such as this can help researchers to implement proper statistical practice and to know when to seek the help of a statistician.O

Acknowledgments

Authors would like to thank Dr. Kyle Grimes for his help in editing this paper.

© 2016 The Obesity Society

References

1. Li P, Brown AW, Oakes JM, Allison DB. Comment on "Intervention Effects of a School-Based Health Promotion Programme on Obesity Related Behavioural Outcomes". *J Obes* 2015;2015:2.
2. Brown AW, Sievenpiper JL, Kyle TA, Kaiser KA. Communication of randomized controlled trial results must match the study focus. *J Nutr* 2015;145:1027-1029.
3. Brown AW, Li P, Bohan Brown MM, et al. Best (but often forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials. *Am J Clin Nutr* 2015;102:241-248.
4. Levitsky DA, Brown AW, Hansen BC, et al. An unjustified conclusion from self-report-based estimates of energy intake. *Am J Med* 2014;127:e33.
5. Dhurandhar NV, Schoeller D, Brown AW, et al. Energy balance measurement: when something is not better than nothing. *Int J Obes* 2014;39(7):1109–1113.
6. Casazza K, Brown AW, Astrup A, et al. Weighing the evidence of common beliefs in obesity research. *Crit Rev Food Sci Nutr* 2014. DOI:10.1080/10408398.2014.922044.
7. Brown AW, Ioannidis JP, Cope MB, Bier DM, Allison DB. Unscientific beliefs about scientific topics in nutrition. *Adv Nutr* 2014;5:563-565.
8. Brown AW, Hall KD, Thomas D, Dhurandhar NV, Heymsfield SB, Allison DB. Order of magnitude misestimation of weight effects of children's meal policy proposals. *Childhood Obes* 2014;10:542-545.
9. Bohan Brown MM, Brown AW, Allison DB. Linear extrapolation results in erroneous overestimation of plausible stressor-related yearly weight changes. *Biol Psychiatry* 2015;78(4):e10–1.
10. Casazza K, Fontaine KR, Astrup A, et al. Myths, presumptions, and facts about obesity. *N Engl J Med* 2013;368:446-454.
11. Brown AW, Bohan Brown MM, Allison DB. Belief beyond the evidence: using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. *Am J Clin Nutr* 2013;98:1298-1308.
12. Brown AW, Allison DB. Unintended consequences of obesity-targeted health policy. *Virtual Mentor* 2013;15:339-346.
13. Schoeller D, Archer E, Dawson JA, Heymsfield S. Implausible results from the use of invalid methods. *J Nutr* 2015;145:150.
14. Li P, Brown AW, Oakes JM, Allison DB. School-based obesity prevention intervention in chilean children: effective in controlling, but not reducing obesity. *J Obes* 2015. DOI:10.1155/2015/183528.
15. Lewis DW Jr, Fields DA, Allison DB. Inconsistencies and inaccuracies in reporting on choice of endpoints and of statistical results in RCT of maternal diet. *Pediatr Obes* 2015. DOI:10.1111/ijpo.12030.
16. Brown AW, Sievenpiper JL, Kyle TA, Kaiser KA, Communication of Randomized Controlled Trial Results Must Match the Study Focus. *Journal of Nutrition* 2015; 145:1027–1029.
17. O'Neill D, Sweetman O. The consequences of measurement error when estimating the impact of obesity on income. *IZA J Labor Econ* 2013. DOI:10.1186/2193-0000-0000-0000.
18. Lyons R. The spread of evidence-poor medicine via flawed social-network analysis. *Stat Politics Policy* 2011. DOI:10.2202/2151-0000.1024.
19. Wake M, Lycett K. Let's call it as it is: on results, reach, and resolution in population-based obesity trials. *Pediatrics* 2014;134:e846-e847.
20. Veličković VM. Opinion: statistical misconceptions. *Scientist*. *The Scientist* 2013. Available from: http://www.the-scientist.com/?articles.view/articleNo/36781/title/Opinion–Statistical-Misconceptions/
21. Torres A. That big childhood obesity decline may have been a statistical error. *National Review* 2014. Available from: http://www.nationalreview.com/corner/373624/big-childhood-obesity-decline-may-have-been-statistical-error-alec-torres.
22. Johns DM. Disconnected? *Slate* 2011. Available from: http://www.slate.com/articles/health_and_science/science/2011/07/disconnected.html.
23. Lykken DT. Statistical significance in psychological research. *Psychol Bull* 1968; 70:151-159.
24. Casella G, Berger RL. *Statistical Inference*, 2nd ed. Pacific Grove: Duxbury; 2002.
25. Cohen J. The earth is round ($P < 0.05$). *Am Psychol* 1994;49:997-1003.
26. Schervish MJ. *P* values: what they are and what they are not. *Am Stat* 1996;50:203-206.
27. Fisher RA. *The Design of Experiments*. London: Oliver and Boyd; 1935.
28. Blackwelder W. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345-353.
29. Koopman KE, Caan MW, Nederveen AJ, et al. Hypercaloric diets with increased meal frequency, but not meal size, increase intrahepatic triglycerides: a randomized controlled trial. *Hepatology* 2014;60:545-553.
30. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review. *Am J Clin Nutr* 2013;97:127-134.
31. Cohen J. The concepts of power analysis. In: Cohen J, editor. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Academic Press, Inc.; 1998. pp 1-17.
32. Cassani RS, Fassini PG, Silvah JH, Lima CM, Marchini JS. Impact of weight loss diet associated with flaxseed on inflammatory markers in men with cardiovascular risk factors: a clinical study. *Nutr J* 2015. DOI:10.1186/1475-0000-0000-0000.
33. Dimova RB, Allison DB. Inappropriate statistical method in a parallel-group randomized controlled trial results in unsubstantiated conclusions. *Nutr J*, in press.
34. Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA* 2010;303:2058-2064.
35. Bland JM, Altman DG. Comparisons against baseline within randomized groups are often used and can be highly misleading. *Trials* 2011;12:1-7.
36. Allison DB, Gorman BS, Primavera LH. The most common questions asked of statistical consultants: our favorite responses and recommended readings. *Genet Soc Gen Psychol Monogr* 1993;119:153-185.
37. Huck SW, McLean RA. Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. *Psychol Bull* 1975;82: 511-518.
38. Myers JL, Well AD. *Research Design and Statistical Analysis*, 1st ed. New York: HarperCollins; 1991.
39. Allison DB. When is it worth measuring a covariate in a randomized clinical trial? *J Consult Clin Psychol* 1995;63:339-343.
40. Albert PS. Tutorial in biostatistics: longitudinal data analysis (repeated measures) in clinical trials. *Stat Med* 1999;18:1707-1732.
41. Kirk RE. *Experimental Design: Procedures for the Behavioral Sciences*, 2nd ed. Pacific Grove: Brooks/Cole; 1982.
42. Liu S, Rovine MJ, Molenaar PC. Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychol Methods* 2012;17:15-30.
43. Shaffer JP. Multiple hypothesis testing. *Ann Rev Psychol* 1995;46:561-584.
44. Mehta TS, Zakharkin SO, Gadbury GL, Allison DB. Epistemological issues in omics and high-dimensional biology: give the people what they want. *Physiol Genom* 2006;28:24-32.
45. Mudge JF, Baker LF, Edge CB, Houlahan JE. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One* 2012;7:e32734. DOI:10.1371/journal.pone.0032734.
46. Bland JPL, Chen L, Cui X, et al. An adaptive alpha spending algorithm improves the power of statistical inference in microarray data analysis. *Bioinformation* 2007; 1:384-389.
47. Elliott TR. Registering randomized clinical trials and the case for CONSORT. *Exp Clin Psychopharmacol* 2007;15:511-518.
48. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med* 2010;7:e1000251.
49. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Bull World Health Organ* 2007;85:867-872.
50. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *CMAJ* 2010;182:1638-1642.
51. Mills JL. Data torturing. *N Engl J Med* 1993;329:1196-1199.
52. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. *PLoS One* 2015;13:e1002106. DOI:10.1371/journal.pbio.1002106.
53. Motulsky HJ. Common misconceptions about data analysis and statistics. *Pharmacol Res Perspect* 2015;3:e00093. DOI:10.1002/prp2.93.

54. Muller K, Fetterman B. *Regression and ANOVA an Integrated Approach Using SAS Software*. Cary: SAS Institute; 2002.

55. Ivanescu AE, Li P, George B, et al. The importance of prediction model validation and assessment in obesity and nutrition research. *Int J Obesity* 2015. DOI:10.1038/ijo.2015.214.

56. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Stat Med* 2007;26:2-19.

57. Hannan PJ. Experimental social epidemiology: controlled community trials. In: Oakes JM, Kaufman JS, editors. *Methods in Social Epidemiology*. San Francisco: Jossey-Bass/Wiley; 2006. pp 335-364.

58. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. *Int Stat Rev* 2009;77:378-394.

59. Feng ZF, Diehr P, Peterson AV, McLerran D. Selected statistical issues in group randomized trials. *Annu Rev Public Health* 2001;22:167-187.

60. Koziol JA, Reid N. On the asymptotic equivalence of two ranking methods for K-sample linear rank statistics. *Ann Stat* 1977;5:1099-1106.

61. Agresti A, Pendergast J. Comparing mean ranks for repeated measures data. *Commun Stat TheoryMethod* 1986;15:1417-1433.

62. Beasley TM. Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behav Res* 2002;37:197-226.

63. Datta S, Satten GA. Rank-sum tests for clustered data. *J Am Stat Assoc* 2005;100: 908-915.

64. Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks*. New York: Springer; 1975.

65. Cliff N. Dominance statistics: ordinal analyses to answer ordinal questions. *Psychol Bull* 1993;114:494-509.

66. Vargha A, Delaney HD. The Kruskal–Wallis test and stochastic homogeneity. *J Educ Behav Stat* 1998;23:170–192.

67. Marascuilo LA, McSweeney M. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Monterey: Brooks-Cole; 1977.

68. Zimmerman DW. A note on homogeneity of variance of scores and ranks. *J Exp Educ* 1996;64:351-362.

69. Zimmerman DW, Zumbo BD. Parametric alternatives to the student *t* test under violation of normality and homogeneity of variance. *Percept Mot Skills* 1992;74: 835-844.

70. Sawilowsky S, Blair RC. A more realistic look at the robustness and type II error properties of the t-test to departures from population normality. *Psychol Bull* 1992; 111:353-360.

71. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials, Panel on Handling Missing Data in Clinical Trials, Committee on National Statistics, Division of Behavioral and Social Sciences and Education*. Washington, DC: National Academies Press; 2010.

72. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am J Epidemiol* 1995;142:1255-1264.

73. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.

74. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;30:377-399.

75. Carpenter JR, Kenward MG. *Multiple Imputation and its Application*. Hoboken: Wiley; 2013.

76. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Converting among effect sizes. *Introduction to Meta-Analysis*. West Sussex: Wiley; 2009, pp 45-49.

77. Gorman BS, Primavera LH, Allison DB. POWPAL: a program for estimating effect sizes, statistical power, and sample sizes. *Educ Psychol Measure* 1995;55:773-776.

78. Bonett DG. Transforming odds ratios into correlations for meta-analytic research. *Am Psychol* 2007;62:254-255.

79. Gøtzsche PC, Hróbjartsson A, Marić K, Tendal B. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* 2007;298:430-437.

80. Kaiser KA, Brown AW, Allison DB. Comment on PMID 25168465: Systematic Review and Meta-Analysis of the Effect of Increased Vegetable and Fruit Consumption on Body Weight and Energy Intake. Bethesda (MD): National Library of Medicine; 2015.

81. Durant N, Baskin ML, Thomas O, Allison DB. School-based obesity treatment and prevention programs: all in all, just another brick in the wall? *Int J Obesity* 2008; 32:1747-1751.

82. Pigott TD. Handling missing data. In: Cooper H HL, Valentine JC, editors. *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation; 2009. pp 399-416.

83. Allison DB, Faith MS. Hypnosis as an adjunct to cognitive-behavioral psychotherapy for obesity: a meta-analytic reappraisal. *J Consult Clin Psychol* 1996;64:513-516.

84. Perry P. Realities of the effect size calculation process: considerations for beginning meta-analysts. In: Bukoski W, editor. *Meta-analysis of Drug Abuse Prevention Programs*. Derby: Diane Publishing; 1997. pp 120-129.

85. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. *Pharm Ther* 2008;33:700-711.

86. Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychol Methods* 2009;14:43-53.

87. Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods* 2002;7:105-125.

88. Hedges LV. Effect sizes in cluster-randomized designs. *J Educ Behav Stat* 2007;32: 341-370.

89. Allison DB, Keating K, Kaiser K, Shikany J. Re: dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *BMJ* 2013;346:e7492. DOI: http://dx.doi.org/10.1136/bmj.e7492.

90. Galton F. Regression towards mediocrity in hereditary stature. *J Anthropol Inst Great Britain Ireland* 1886;15:246-263.

91. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol* 2005;34:215-220.

92. Chobanian AV, Bakris GL, Black HR, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension* 2003;42:1206-1252.

93. Senn S. Francis Galton and regression to the mean. *Significance* 2011;8:124-126.

94. Cadzow RB, Chambers MK, Sandell AM. School-based obesity intervention associated with three year decrease in student weight status in a low-income school district. *J Community Health* 2015;40:709-713.

95. Sandercock GR, Cohen DD, Griffin M. Evaluation of a multicomponent intervention to improve weight status and fitness in children: upstarts. *Pediatr Int* 2012;54:911-917.

96. King MH, Lederer AM, Sovinski D, et al. Implementation and evaluation of the HEROES initiative a tri-state coordinated school health program to reduce childhood obesity. *Health Promot Pract* 2014;15:395-405.

97. Kim N, Seo DC, King MH, Lederer AM, Sovinski D. Long-term predictors of blood pressure among adolescents during an 18-month school-based obesity prevention intervention. *J Adolesc Health* 2014;55:521-527.

98. Castro DC, Samuels M, Harman AE. Growing healthy kids: a community garden-based obesity prevention program. *Am J Prev Med* 2013;44:S193-S199.

99. Skinner AC, Heymsfield SB, Pietrobelli A, Faith MS, Allison DB. Ignoring regression to the mean leads to unsupported conclusion about obesity. *Int J Behav Nutr Phys Activity* 2015;12:56.

100. Allison DB, Loebel AD, Lombardo I, Romano SJ, Siu CO. Understanding the relationship between baseline BMI and subsequent weight change in antipsychotic trials: effect modification or regression to the mean? *Psychiatry Res* 2009;170:172-176.

101. Yudkin P, Stratton I. How to deal with regression to the mean in intervention studies. *Lancet* 1996;347:241-243.

102. Cook TD, Campbell DT, Day A. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin; 1979.

103. Smith R. David Sackett Physician, trialist, and teacher. *BMJ* 2015;350:h2639. DOI:10.1136/bmj.h2639.

104. Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51-63.

105. Collins FS, Tabak LA. NIH plans to enhance reproducibility. *Nature* 2014;505: 612-613.

106. Popper K. *The Logic of Scientific Discovery*. New York: Routledge; 2002.

107. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011;22:1359-1366.

108. Gelman A. "False-positive psychology". Available from: http://andrewgelman.com/2012/02/16/false-positive-psychology/.

109. Gelman A. Too Good to Be True. *Slate* 2013. Available from: http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html

110. Hayes AF. Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Commun Monogr* 2009;76:408-420.

111. Moher D, Liberati A, Tetzlaff J, Altman DG. The Prisma Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.