Published in final edited form as:
Obes Rev. 2019 November ; 20(11): 1523–1541. doi:10.1111/obr.12923.

Childhood Obesity Intervention Studies: A Narrative Review and Guide for Investigators, Authors, Editors, Reviewers, Journalists, and Readers to Guard Against Exaggerated Effectiveness Claims

Andrew W. Brown,
Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, IN, 47405, USA

Douglas G. Altman[+],
Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

Tom Baranowski,
Department of Pediatrics, Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, TX, 77030

J. Martin Bland,
Department of Health Sciences, University of York, York, United Kingdom

John A. Dawson,

Corresponding authors Andrew W Brown, PhD, SPH 116, 1025 E. Seventh Street, Bloomington, IN, 47405, awb1@iu.edu, David B Allison, PhD, SPH 111, 1025 E. Seventh Street, Bloomington, IN, 47405, allison@iu.edu.

[+]Prof. Altman contributed to this article prior to his untimely passing in June of 2018. His inclusion as an author here recognizes his contributions, though he was unable to approve of the final version.

Disclosures

Dr. Allison has received personal payments or promises for same from: Biofortis; Fish & Richardson, P.C.; HawkPartners; IKEA; Laura and John Arnold Foundation; Law Offices of Ronald Marron; Sage Publishing; Tomasik, Kotin & Kasserman LLC; Nestec (Nestlé); WW (formerly Weight Watchers International LLC). Donations to a foundation have been made on his behalf by the Northarvest Bean Growers Association and the United Soy Bean Board. Dr. Allison is an unpaid member of the International Life Sciences Institute North America Board of Trustees. Dr. Allison's institution, Indiana University, has received funds to support his research or educational activities from: Alliance for Potato Research and Education; Dairy Management Inc.; Herbalife; and the NIH. His prior institution, the University of Alabama at Birmingham, received grants, gifts or contracts from multiple food, beverage, and other for profit and non for profit organizations with interests in obesity, statistical methods, and research design. Dr. Baranowski discloses being employed by Baylor College of Medicine, his institution having NIH grants for his work, and having received speaking fees from the University of Georgia. Dr. Bland discloses that topics presented herein are related to a textbook for which he receives royalties; and has received travel accommodations from the University of Western Ontario and for the Health Services Research Board Senior Investigator Award. Dr. Brown has received travel expenses from Academy of Nutrition and Dietetics, National Academy of Sciences, and University of Michigan; speaking fees from Academy of Nutrition and Dietetics, American Society for Nutrition, Birmingham District Dietetic Association, Kentuckiana Health Collaborative, and Rippe Lifestyle Institute, Inc.; and grants through his institution from Dairy Management, Inc. and the NIH. He has been involved in research for which Indiana University, the University of Alabama at Birmingham (his past institution) or colleagues have received grants or contracts with Communiqué, the Sloan Foundation, and multiple food, beverage, and other for profit and non for profit organizations with interests in obesity, statistical methods, and research design. Dr. Dawson discloses his employment by Texas Tech University, grants from the Egg Nutrition Center/ American Egg Board, and travel expenses from the American Society for Nutrition. Dr. Dhurandhar discloses serving as Editor-In-Chief of *Nutrition and Diabetes*, lecture fees from Metabologix, and manuscript preparation fees from WebMD. Dr. Fontained discloses stock in Virta Health as an advisory board member, and payment as an advisory board member for Atkins Nutritionals. Dr. Heymsfield discloses paid service as on medical advisory boards of Medifast Corp and Tanita Corp. Mr. Kyle discloses consultancy with Nutrisystem and Novo Nordisk. Dr. Stevens discloses grants with NIH and WW through her institution. Drs. Dowla, Gelman, Jayawardene, Keith, Loken, Oakes, and Thomas have no disclosures.

Department of Nutritional Sciences, Texas Tech University, Lubbock, TX, 79409

Nikhil V. Dhurandhar,
Department of Nutritional Sciences, Texas Tech University, Lubbock, Texas 79409

Shima Dowla,
School of Medicine, University of Alabama at Birmingham, Birmingham, AL, 35294, USA

Kevin R. Fontaine,
Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Andrew Gelman,
Department of Statistics and Department of Political Science, Columbia University, New York

Steven B. Heymsfield,
Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA

Wasantha Jayawardene,
Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, IN, 47405, USA

Scott W. Keith,
Department of Pharmacology and Experimental Therapeutics, Division of Biostatistics, Sidney Kimmel Medical College, Thomas Jefferson University, 1015 Chestnut St., Suite 520, Philadelphia, PA, 19107, USA

Theodore K. Kyle,
ConscienHealth, Pittsburgh, PA

Eric Loken,
Neag School of Education, University of Connecticut, Storrs, CT

J. Michael Oakes,
Department of Epidemiology, School of Public Health, University of Minnesota, 1300 South 2nd St, Minneapolis, MN 55454

June Stevens,
Departments of Nutrition and Epidemiology, Gillings School of Global Public Health, University of North Carolina, CB7400, Chapel Hill, NC 29599

Diana M. Thomas,
Department of Mathematical Sciences, United States Military Academy, West Point, NY

David B. Allison
Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, 47405, USA

Abstract

Being able to draw accurate conclusions from childhood obesity trials is important to make advances in reversing the obesity epidemic. However, obesity research sometimes is not conducted or reported to appropriate scientific standards. To constructively draw attention to this issue, we present 10 errors that are commonly committed, illustrate each error with examples from the childhood obesity literature, and follow with suggestions on how to avoid these errors. These errors are: Using self-reported outcomes and teaching to the test; Foregoing control groups and risking regression to the mean creating differences over time; Changing the goal posts; Ignoring clustering in studies that randomize groups of children; Following the forking paths, sub-setting, p-hacking, and data dredging; Basing conclusions on tests for significant differences from baseline; Equating 'no statistically significant difference' with 'equally effective'; Ignoring intervention study results in favor of observational analyses; Using one-sided testing for statistical significance; and, Stating that effects are clinically significant even though they are not statistically significant. We hope that compiling these errors in one article will serve as the beginning of a checklist to support fidelity in conducting, analyzing, and reporting childhood obesity research.

Keywords

Childhood obesity; causal inference; interventions

Introduction

> "Experimental scientists must have for data a permanent respect that transcends their passing interest in the stories they make up about their data."[1] Cletus J. Burke, 1954

Childhood obesity is a substantial global public health concern that, despite many efforts, has continued to climb for decades,[2] and few would argue with the merit of pursuing effective prevention or treatment options. Substantial resources are dedicated to studying childhood obesity.[3] However, when prevention or treatment programs use popular or seemingly wholesome practices based on cherished principles, some people might believe that questioning the merits of such programs is inappropriate, or even that doing so subverts or undermines public support for implementing and funding such interventions. Yet, society must increasingly ask whether proposed solutions are evidence-based. Thus, unvarnished presentations of evidence regarding the effectiveness of such programs is vital. Nevertheless, the extent to which studies that assess obesity interventions demonstrate effectiveness of the interventions has been substantially overstated in some cases, leading to concerns about the rigor of childhood nutrition and obesity research in particular.[4] This observation is not based on a systematic quantification, yet illustrative cases are easy to find when reading the literature from countries around the world. At the very least, such cases demonstrate there is room for improvement.

The scientific community, and those who rely on the community's work, need accurate information for informed conclusion- and decision-making. Therefore, we delineate 10 errors that exaggerate the apparent extent to which interventions lead to positive improvements in obesity-related outcomes, with a focus on examples from the childhood obesity literature. We use the word 'intervention' to include programs, policies, or prescriptions to treat or prevent obesity and obesity-related outcomes. Errors may apply to both controlled and uncontrolled studies; or to randomized and non-randomized experiments. We describe these errors, supported by examples in published studies, and make recommendations to avoid them.

Our use of specific examples is not meant to impugn specific researchers, make judgments of intentionality, or make conclusions about the ultimate effectiveness of interventions. In some examples throughout, the underlying data and interventions appear sound, and analytic or communication errors could explain the discrepancy. One recent case has called into question multiple publications, resulting in multiple obesity-related papers (some related to childhood obesity) being retracted (c.f., six retractions in one notice[5]). Herein, we point out that the published errors exist; any errors in the literature weaken the evidence base regardless of intentionality. We also note these errors are not necessarily limited to the field of childhood obesity; some of these or related errors have been identified in the field of maternal and child nutrition,[6] in obesity research more generally,[7] and in science more broadly[8]. Finally, this list is not exhaustive, and the order of presentation herein does not imply ranking, prioritization, or severity among the errors.

We hope this article can serve as a partial checklist of mistakes to be avoided. By highlighting the errors here, authors may be better able to avoid them, and reviewers, editors, journalists, and other readers will be better able to detect the mistakes and adjust their conclusions and actions accordingly.

Inferential Error: Using Self-Reported Outcomes and Teaching to the Test

Error Description

Implement a program that urges the intervention group to change health-related behaviors or conditions, and then give participants a questionnaire that asks about the same health related behaviors and conditions, ignoring the differential bias this practice can induce.[9]

Explanation of the Error and Why the Practice is Wrong

As a simple example, teaching to the test in a childhood obesity intervention could be to encourage children to eat more of a healthy food (the teaching), and considering the children compliant when they report eating more of that food (the test), whether or not they actually do. Stated another way, bias induced by an intervention is a type of social desirability bias (i.e., the tendency for individuals to answer or portray themselves in such a way to avoid criticism, adhere to perceived cultural norms, or garner praise).[10] This can be a particular concern for studies of youth, because school-aged children may be especially prone to "report more socially desirable behavior (or less socially undesirable behavior) when they fear that this information is shared with their parents or other adult authorities."[11] In the context of an intervention, social desirability bias can be stronger or manifest differently in the intervention group because, by the nature of the intervention, those individuals have been coached to change the behaviors that they are subsequently asked about. A few studies have compared the discrepancy between self-reported and objectively measured data in participants in intervention versus control groups. Intervention-induced bias in self-reported diet, physical activity, and body weight outcomes was detected in some[12–15] but not all[16,17] studies. In one study that did not detect bias, the investigators took special care to separate the data collection from the intervention, using three different teams of staff and deceiving the subjects that the goal of the study for which the data were collected was different from the actual goal of the intervention.[16]

Examples of the Error

Most weight control interventions use measured rather than self-reported body weight as the primary outcome, but self-report has been used. Self-report measures are used more often to assess intervention effects on physical activity and almost always for diet. Several studies have described differences in self-reported intake[18–20] and/or physical activity[18,21,22] between the intervention and control groups despite no impact of the intervention on measured BMI or body weight. In one illustrative case, investigators implemented an intervention to promote physical activity. Compared to the control, the intervention group self-reported greater physical activity, but the objective accelerometry data did not detect a difference between groups.[23] When the self-reported measures are used, authors often indicate measurement error as a limitation,[18–21] but rarely mention the possibility of intervention-induced bias.[18]

Recommendations

Since intervention-induced bias exists in some studies, and because the face validity for its potential is strong, we discourage the use of self-report in trials when feasible objective measures exist, such as body weight and physical activity. For dietary intake (a key component in most weight-related interventions), objective methods are not readily available in most studies. In those circumstances, we advise investigators to forego emphasizing intervention effects on self-reported energy consumption in particular,[24,25] and to remind the reader that bias related to the intervention can occur when diet is measured by self-report. Additionally, we suggest that the term “self-report” be specifically mentioned in the abstract if data are self-reported.

Self-report biases are likely to be found in the same types of individuals who show other types of social desirability bias.[26] Research on the efficacy of strategies to reduce the perceived link between the self-reported information and the intervention could result in methods to reduce bias and improve data quality. More research on the attributes of self-report biases in studies that include weight-related interventions is merited.

Inferential Error: Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time

Error Description

Provide an intervention to a sample that consists entirely of individuals greater or less than the average on some characteristic – such as children all with high BMI z-scores – with no control group and assume improvements in the variable result from the intervention, rather than a spontaneous tendency for extreme values to revert toward the population average.

Explanation of the Error and Why the Practice is Wrong

In 1886, when evaluating offspring height relative to tall parents, Sir Francis Galton observed the phenomenon he initially referred to as "regression toward mediocrity."[27] Specifically, and perhaps surprisingly, Galton found offspring were shorter than their parents if they had tall parents. He recognized that by first considering a portion of the population that holds extreme values (e.g., tall individuals), the second measurement (e.g., their offspring's height) would be closer to the population average and hence the offspring would be shorter. Later, Galton revised the name of this observation to what we today know as regression to the mean (RTM), with much written about examples and methods to avoid or address it over the years (c.f.,[28,29]).

Unfortunately, childhood obesity investigators sometimes erroneously conclude positive effects of an intervention that can be attributed to RTM. This typically occurs when a population with extreme baseline values is investigated, such as children with high BMI z-scores (BMIz; a child's BMI standardized to a reference distribution, such as those proposed by the International Obesity Task Force[30]). In some cases, investigators exacerbate this phenomenon by analyzing the data by subgroups of baseline levels. When the group is re-measured at the end of the study, the score is lower, with investigators drawing the conclusion that the intervention was effective. However, as observed by Galton, by RTM alone, we expect an extreme group to have lower values at a subsequent point in time. We clarify that RTM does not imply that children with BMIz in the obesity range are expected to spontaneously revert to the normal BMIz range, which would be truly remarkable. Rather, in RTM the subsequent measurements are expected to be lower on average; how much lower depends on many factors related to measurement error, natural variability, and the extremeness of the selected subgroup.

Examples of the Error

A holistic health intervention designed to improve knowledge of and employ healthful behaviors was implemented in 40 participating elementary schools.[31] BMI-for-Age z scores were recorded at baseline and the authors concluded program effectiveness due to the largest decreases of BMIz at the end of the school year in students who were classified with overweight or obesity at baseline. Using the 1997 National Longitudinal Survey of Youth data as a benchmark strongly supported that these decreases were not a result of the intervention but were attributable to RTM.[32] Similarly, when evaluating the impact of nutrition education on African American preschoolers,[33] study authors concluded positive intervention effects when considering only children in the intervention group with overweight or obesity. When the possibility of RTM was suggested to the authors,[34] they tested and found a decrease in the control group BMI consistent with RTM, and should be commended for publishing a clear correction that stated, "we cannot make any affirmative statements about the effectiveness of our interventions."[35] Finally, a physical activity intervention program[36] that enrolled only children with overweight or obesity found a decrease in BMIz at the post-intervention measurement, again consistent with RTM. More examples exist (e.g.,[37,38]).

Recommendations

The best practice to determine a true intervention effect is to include a control group from the same population because RTM will impact the control group as well as the intervention group under standard assumptions (e.g., no bias from differential attrition between the two groups).[39] If a control group is not included, the effects of RTM can still be estimated by predicting the expected second measurement from knowledge of the measurement's reliability and the population mean.[40] Multiple baseline measurements could also help inform the potential degree of RTM effect. At the very least, authors should clearly, and without reservation, acknowledge the distinct possibility that RTM could explain the improvements after intervention. Watson et al. did just that when communicating their results on a family-based childhood obesity program, albeit without reference to RTM by name: "As with many service evaluations, this study is limited by a lack of control group and a high attrition rate. It is not therefore known what change might have occurred without intervention."[41]

Inferential Error: Changing the Goal Posts

Error Description

When a study to test an intervention's effect on obesity yields a non-significant result for the primary outcome, use surrogate secondary outcomes to make claims of effectiveness for an intervention.

Explanation of the Error and Why the Practice is Wrong

A meta-analysis reported that 79% of interventions to prevent or reduce childhood obesity were unsuccessful.[42] Interventions failing to show an effect are therefore the norm. Yet, rather than reporting a non-significant result for the primary outcome of childhood obesity interventions (e.g., BMIz, body weight), some investigators emphasize or only report success based upon secondary outcomes for surrogate obesity measurements or presumed intermediate drivers of obesity such as increased knowledge, improved attitudes, reduced self-reported dietary intake, or increased physical activity. In one version of this error, authors may conclude that success in altering surrogate outcomes support an intervention's use for improving obesity, despite no improvements in obesity; in another, the authors may ignore the original primary goal of affecting obesity, and instead make conclusions about the surrogate outcomes alone. Often the reader is not informed of the original primary goal. This technique of changing the criteria for success is commonly referred to as changing or moving the goal posts.[43]

While nutrition-related knowledge or behavior, intake of energy or various nutrients, physical activity, and many other factors may be intermediate drivers of BMI or obesity, it is unreasonable to use them as surrogate markers for obesity itself. A major downside to changing the goal posts is that interventions are reported as effective even though they did not satisfy the pre-specified objective: to prevent or treat childhood obesity. Advocacy for such ineffective interventions as strategies for combating childhood obesity is then added to the literature, giving the false appearance of an increasing body of supporting evidence that yields confidence in efficacy of intervention approaches that, in fact, were not successful.

Examples of the Error

The stated aim of a cluster randomized controlled trial of nutrition education was to use social cognitive theory (SCT) to reduce and prevent obesity among adolescent girls. The study concluded: “Although school-based nutrition education intervention using SCT did not change significantly BMI and WC among the targeted population in this study, dietary habits as well as psychological factors improved significantly in the intervention group.”[44] Although the study did not affect the stated aim of obesity outcomes, the authors still concluded that a “school-based intervention based on SCT introduces a new approach to health authorities” based on surrogate measures.

Another study using a school-based, cluster-randomized design implemented health-promoting strategies for 3.5 years.[45] There were no significant differences between the control and intervention group for the majority of the stated primary and secondary outcomes, including BMI, BMIz, and prevalence of overweight and obesity. The authors admitted that “only limited translation of those environmental changes into improved behaviours and weight status were evident at follow up.” Yet, they concluded that, “[t]his 3.5 year intervention demonstrates that it is possible to effect system level change and some improvements in health and wellbeing outcomes from investments that focus on the school environment…” In addition, despite no statistical significance, they declared changes in outcome variables such as vegetable consumption, as a positive outcome.

In the above examples, no effects on obesity were demonstrated and so the outcome focus changed to general statements of health. Sometimes, no effect will be seen on obesity, yet promising results in a surrogate outcome may lead authors to still conclude effects on obesity. As described in a letter to the editor in one such case, the original researchers saw no statistically significant differences in their primary obesity measurement, demonstrated only a single statistically significant difference among a battery of non-registered anthropometric measurements, and still concluded that their intervention may benefit infant adiposity.[46]

Recommendations

Authors of intervention studies to reduce childhood obesity should clearly indicate the results pertinent to the pre-specified primary hypothesis and not obscure those findings by excessive focus on alternative outcomes. We are not discouraging the collection, analysis, or reporting of secondary or surrogate endpoints, but it is important that the primary outcomes are decided in advance and communicated clearly and completely, and alternative endpoints are distinguished appropriately.[47]

A study by Lloyd et al.[48] offers an exemplary approach for drawing conclusions. This obesity prevention trial of children from 32 schools observed no significant effect on obesity. The authors concluded, “we found no effect of the intervention on preventing overweight or obesity. Although schools are an ideal setting in which to deliver population-based interventions, school-based interventions might not be sufficiently intense to affect both the school and the family environment, and hence the weight status of children”. Importantly, the study did not advocate for the repeat of the same approach. Instead, it recommended that "[f]uture research should focus on more upstream determinants of obesity and use whole-systems approaches." Similarly, Barkin et al.[49] noted that their preschool-age intervention did not significantly affect BMI trajectories over 36 months, but did find a significant difference in reported energy intake in favor of the intervention (see "Using Self-Reported Outcomes and Teaching to the Test"]. Nevertheless, the abstract remained focused on the primary outcome: "A 36-month multicomponent behavioral intervention did not change BMI trajectory … compared with a control program. Whether there would be effectiveness for other types of behavioral interventions or implementation in other cities would require further research."

Journal editors and reviewers should encourage publishing all well-conducted studies, including results from interventional strategies that did not improve childhood obesity. This may ease pressure on authors to provide "spin" [50,51] on an interventional study with null findings. Where spin exists, it needs to be corrected by reviewers and editors before publication. Finally, readers need to be sure to skeptically read and interpret results.

Inferential Error: Ignoring Clustering in Studies that Randomize Groups of Children

Error Description

Conduct a cluster randomized trial, in which groups of children (e.g., entire classrooms, schools, or pediatric clinics) are randomly assigned to experimental conditions, but analyze the data as though the children were randomized individually.

Explanation of the Error and Why the Practice is Wrong

There are two key aspects to this error, clustering and nesting, and ignoring either can weaken or invalidate statistical inference and thus conclusions. With respect to clustering: children from in-tact social groups, such as classrooms, clinics, or even neighborhoods, tend to be more highly correlated within a cluster than between clusters. In simplest terms, children in one classroom may tend to be more alike than children in another classroom. Reasons for this may include social selection (e.g., educational tracking or impacts of efforts to maintain friendship networks) and common exposures (e.g., teacher A versus teacher B). Statistically, this means that we have less *independent* information than we would expect from a simple, random, non-clustered sample. Less information means the effective sample size is less than the actual, or nominal, sample size. For example, there may be 100 children in a study but as a result of clustering the study may have information equivalent to only 80 independent children.[52] Classical regression methods, like ordinary least squares or logistic regression, and classical hypothesis tests, like Student's t-test or Pearson's chi-square test, are predicated on the observations being statistically independent. Applying these classical estimation and inference methods to correlated observations from cluster-randomized trials tends to underestimate standard errors, which erroneously makes p-values smaller than they should be, and increases the risk of falsely rejecting a null hypothesis of no intervention effect (i.e., making a type I error).[7] Simply put, analyses that ignore clustering may yield smaller p-values than proper analyses that incorporate clustering. The Consolidated Standards of Reporting Trials (CONSORT) extension for Cluster Trials, which are best-practice reporting guidelines for cluster trials, include the advice that cluster randomized trials “should not be analysed as if the trial was individually randomized...”[53] The issue was further highlighted by the National Institutes of Health in their “Research Methods Resources” website: “Any analysis that ignores the extra variation … or the limited [degrees of freedom] will have a type 1 error rate that is inflated, often badly.”[54] Thus, ignoring clustering risks type I errors (i.e., concluding there is a difference between groups when a difference does not exist). On the other hand, ignoring the correlated observations in the planning stages of a cluster randomized trial means that cluster randomized trials may be underpowered when analyzed correctly and thus researchers risk making type II errors, as well (i.e., failing to conclude there is a difference between groups when a difference actually exists).[55]

The second issue is nesting, which is to say the randomized clusters (e.g., schools) are nested or wholly located within experimental conditions. As a result, the unique aspects of the clusters themselves (e.g., percentage receiving free and reduced lunch, age of school building, or tax-base supporting the school) may confound intervention effects. To eliminate the threat of such cluster-specific confounding from desired intervention effects, one must have many replicate clusters within experimental conditions. Such cluster-level replicates determine the degrees of freedom (which, roughly speaking, represent the amount of independent information) available for testing intervention effects. Thus, studies cannot have just one cluster per experimental condition: doing so yields zero degrees of freedom for intervention effects. The CONSORT extension for Cluster Trials summarizes the problem by noting “[t]rials with one cluster per arm should be avoided as they cannot give a valid analysis, as the intervention effect is completely confounded with the cluster effect.”[53]

Though we focus on groups of children, these concerns apply just as much to groups of parents, teachers, or others targeted by an intervention intended to address childhood obesity.

Examples of the Error

Many existing studies randomized clusters of subjects to study groups but subsequently ignored the clustering in the statistical analyses (as reviewed in[56] and addressed in letters to editors[57,58]). In one example,[44] researchers evaluated anthropometric, nutritional behavior, and social cognitive outcomes among 173 adolescent girls with overweight or obesity assigned to either an intervention or control group. Despite the authors following published guidelines on reporting cluster randomized trials,[53] their analyses did not account for the fact that the girls were students belonging to one of 8 schools randomized to the intervention or control groups. Even if the intra-cluster correlation in the observations within these schools were as low as 0.05, the variance inflation[59] caused by ignoring the clustering would be at least 2.03 under reasonable assumptions, suggesting that their reported outcome variance estimates are likely at most half of the unbiased outcome variance estimates corrected for the clustering. This might have had a profound, invalidating impact on inferences made in that study.

Some examples involve using too few clusters. In one such example, authors included two schools in each of two districts to estimate the effects of a multi-component, school-based intervention.[60] A letter expressing concerns about this paper[61] noted that despite the authors recognizing the importance of including clustering *a priori*, the authors failed to include clustering in analyses, and even compared pairs of schools within districts, resulting in tests that would have had zero degrees of freedom if analyzed correctly (i.e., there would be no information to estimate the variability in differences between the groups). In response, the study authors justified their use of incorrect analyses in part by citing others who also used too few clusters,[62] reinforcing the importance of preventing such misanalysed studies from appearing in the literature to begin with. In response to a subsequent critique of the same study,[63] the original authors published a corrigendum that continued to make invalid causal conclusions about their intervention.[64] In another case, investigators randomized one school each to 4 interventions, plus 4 no-intervention control schools.[65] A critique of the study noted that "although the number of clusters that are needed in a cluster-randomized trial is not fixed, that number is never 1," and therefore that the study "could not establish causation and, at best, only had the capacity to create the hypothesis that [the interventions] may have a favorable impact on childhood obesity."[66] In one other case, an article making similar mistakes was retracted "because the statistical analysis was not correct given the cluster-randomized design" and the "conclusion that the original paper drew about having demonstrated treatment efficacy was not supported in the corrected analysis."[67]

Recommendations

The degree to which these issues impact the validity of a cluster randomized trial depends on many things, perhaps most notably the number of clusters randomized, the number of children in each cluster, and how highly correlated the observations are within clusters (i.e., the intra-cluster correlation which can be measured by the ratio of the between-cluster outcome variance to the total outcome variance).[59] These and other fundamental issues with cluster randomized trials and modern practices for addressing the issues have been described in detail elsewhere[56] and thorough reviews of design and analysis methodologies for these trials were recently published.[68,69] A rule of thumb is that studies should have at least 10 clusters per experimental condition to have a chance of reasonable power to detect large intervention effects, and such tests must rely on the t-distribution, which adjusts for the limited sample size. Thirty or more clusters per experimental condition are needed for z-tests of intervention effects (i.e., the normal approximation to the t-distribution for large samples). Power analyses and statistical analyses need to include the clustering to appropriately control expected type I and type II errors. In the case of single clusters per group, authors need to be explicit about the downgrading of the study from a cluster-randomized trial to a quasi-experiment because clusters are perfectly confounded with intervention.

Inferential Error: Following the Forking Paths, Sub-Setting, P-Hacking, and Data Dredging

Error Description

If results are not statistically significant with the preplanned primary analysis in the total sample, or if there is no preplanned analysis, keep trying different analyses with different subsets of the sample or various outcomes and base conclusions on whatever is statistically significant.

Explanation of the Error and Why the Practice is Wrong

To report an intervention effect with $p < 0.05$ generally means that if the null hypothesis were true, appropriately calculated test statistics that are equal or greater in magnitude to that observed would occur in fewer than 5% of samples.[70] When many possible analyses of the data are performed, and if the null hypothesis is true, the probability of finding at least one statistically significant result by chance increases. Simmons, Nelson, and Simonsohn introduced the phrase "p-hacking" in their demonstration of how flexible stopping rules for recruitment, testing multiple outcomes, and exploring for interaction effects could dramatically raise the chance of a false positive[71]; similar approaches have been referred to as "undisclosed flexibility in data collection," "researcher degrees of freedom,"[72] "data dredging",[73] and "following the forking paths,"[74] among other names, with the authors making nuanced distinctions amongst these terms. Generally speaking, if analytical choices are made based on features of the data at hand rather than *a priori* decisions or pre-specified theory, it is possible that the p-value no longer represents the probability under the null hypothesis, and highlights the importance of preregistering studies and analyses. As a concrete example, consider researchers who decide to pool overweight and obesity into the same category after looking at the data because the number in the obesity category is too small and thus underpowered. Grouping overweight and obesity might be a legitimate decision under some circumstances, but when made after the data are collected and evaluated, it raises the question of whether those categories would have been pooled if the number in the obesity category was larger. It is important to note that the problem arises even when such selection is unintentional, such as many implicit tests for samples that may have been analyzed differently.[75,76]

Examples of the Error

It is often difficult to determine whether inappropriate or undisclosed analytic flexibility occurs in any specific case without knowing *a priori* what authors intended to do. Besides p-curve analyses,[72] the best evidence may come from comparing randomized trials to their preregistrations. As described with "Changing the Goal Posts," discordance between registered primary outcomes and the reporting in manuscripts can reveal analytical or reporting decisions. Discordance between registration documents and publications is not uncommon in obesity literature.[51,77] In addition, flexibility in analyses can be detected even through the number of participants included in analyses. In three studies reported from the ACTIVITAL study,[78] total sample sizes were reported as 1370, 1430, and 1440, and the sample sizes used for analyses included 1046, 1083, and 1224. In addition, one of the papers focused on subgroup analyses.[79] In some cases, subgroups can be important in evaluating the results of a trial [cf.,[80]], particularly when the subgroups are pre-specified. However, subgrouping can also be associated with researchers wandering through the forking paths of research decisions.[74] For instance, the authors categorized students into different activity categories based on accelerometer counts, but did not cite or pre-specify the thresholds. It is therefore unclear if the cutpoints were established *a priori* or based on the data. Conversely, they do cite *a priori* thresholds for subgrouping households by poverty status, fitness group by established standards, and BMI by International Obesity Task Force criteria. In the latter case, however, the authors chose to pool overweight with obesity. The decisions of sample sizes, cutoffs, and pooling of groups may be perfectly legitimate, but the full process of how the decisions were made is unclear from the reports and the registration, and it is uncertain what effect the flexibility of choices may have had on the final results.

Recommendations

Determining whether p-hacking occurred in a single paper can be difficult even with preregistration. However, approaches called p-curve and p-uniform[72,81,82] were developed to evaluate the distribution of many p-values observed across many studies, such as from a meta-analysis, or multiple analyses within a single study, to test for specific patterns in the p-values. Others have introduced text-mining techniques to investigate p-hacking in scientific literature and test for p-hacking when conducting a meta-analysis.[83] Although not perfect, these methods have been used at least once in the childhood obesity and exercise literature.[84] The results suggested that selective reporting was not obviously present, and the authors suggested that the results were not intensely p-hacked from this small subset of studies.

Researchers can protect against inappropriately capitalizing on chance findings in multiple ways. One familiar approach is to correct for multiple comparisons or attempt to control the false discovery rate. These methods control the type I error rate across multiple comparisons, but in so doing make it harder to reject the null hypothesis (i.e., decrease power), and, hopefully, encourage researchers to make fewer and more focused analyses. Nevertheless, p-value adjustments would depend upon a careful counting of all tests conducted, not just those published, and can fast become unwieldy. In addition, researchers can pre-register their analysis plan and main hypotheses. Pre-registration can protect against any appearances that results were obtained through undisclosed p-hacking, and will likely constrain the number of analyses. In some situations, preregistration is required.[85] Alternatively, multiple outcomes can be combined into one analysis using hierarchical modeling,[86] which can mitigate multiple testing concerns. In this way, researchers can present more comparisons of interest and then analyze them together, rather than presenting only fewer or a single pre-chosen comparison (which would limit our ability to learn from data). In any approach, all results should be presented, whether or not the results reach predefined statistical significance thresholds. We do not mean to discourage performing creative or exploratory data analyses. Rather, what is important is openness. Randomized trials should pre-specify primary and secondary outcomes, report the “multiverse” of analyses tried, and describe the analytical paths taken, rather than selecting the subset that achieve some arbitrary threshold of “statistical significance” or desirable results.

Inferential Error: Basing Conclusions on Tests for Significant Differences from Baseline

Error Description

Separately test for significant differences from baseline in the intervention and control groups and if the former is significant and the latter is not, declare the result statistically significant.

Explanation of the Error and Why the Practice is Wrong

Researchers often want to compare the level of a variable between two groups over time. These might be experimental, as in a randomized trial, or observational, as in a cohort study. In both of these designs, we often have an observation of the variable at baseline and follow-up. Some researchers test for changes over time within groups. If one group shows a statistically significant change from baseline, and the other group does not, sometimes authors will conclude that there is a difference between groups. However, no formal between-group test was conducted. This interpretation involves regarding the non-significant difference in one group as showing no difference (i.e., accepting the null), and the significant difference in the other group being interpreted as concluding there is a difference (i.e., rejecting the null). However, "not significant" does ***not*** imply "no difference", only that we do not have sufficient evidence that a difference exists between groups. Testing for differences between groups by separate analyses of within-group changes is also referred to as the *D*ifferences *i*n *N*ominal *S*ignificance (DINS) error[8] or inappropriate testing against baseline values.[7]

It is useful to simulate this method of analysis for the situation in which we know that there is no difference between groups (i.e., the null hypothesis is true). Two of us[87,88] simulated a two group, pre-post design. At the simulated baseline, we generated random observations from the same population, hence having no underlying differences (mean of 0), with a standard deviation of 2.0. We then simulated a random change from baseline to each observation to simulate a follow-up measurement, having the same mean of 0.5 and standard deviation of 1.0. We then carried out paired t tests in each group to test for change from baseline. We found that in 10,000 runs of this simulation, 617 (6.2%) pairs of groups had neither test significant, 5,675 (56.8%) had both tests significant, and 3,708 (37.1%) had one test significant but not the other. Hence, for this particular set-up, where both groups come from the same population and the null hypothesis that the groups come from populations with the same mean is therefore true, the probability of detecting a difference using the separate test strategy is not the 5% we should have, but 37.1%.

If the probability of detecting a statistically significant result for a change over time within each group is P (that is, P is the power to detect a difference over time), the probability that one group will have a significant difference and the other will not is $2P(1-P)$.[88] 2P(1–P) has a maximum value of 0.5 when $P = 0.50$, so that half of all such trials would show a significant difference in one group but not in the other, even if the null hypothesis of no difference between groups is true. If the changes over time also have true null hypotheses, so that there are no differences over time or between groups, the probability of one significant and one not significant comparison of change over time is 2×0.05×(1 – 0.05) = 0.095 – i.e. about twice the nominal 5%. Thus, the separate tests procedure is always misleading.

If the powers for the two tests against baseline are different, P_1 and P_2, the probability of one test being significant and one non-significant becomes $P_1(1 - P_2) + P_2(1 - P_1)$, which can be close to 1 if one power is large and the other small. The differences in P_1 and P_2 can be caused by very different group sizes with identical effect sizes (that is, the null hypothesis is true), or the differences from baseline could vary greatly between groups (that is, the null hypothesis is false). In the latter case, of course, there would be a difference between the groups, but an invalid analysis is still inappropriate, even if it produces the “correct” answer by chance, because in practice we do not know which situation is true.

Statistically significant changes from baseline within a group may be due to the intervention, but there are several other possibilities, including random chance, seasonal variation, systematic changes with age, and regression towards the mean (see “Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time”). We can expect that in a study of obesity, especially in children, the mean height, weight, BMI, or other measurements may change over time and the power of the pre-post test to detect a change may be considerably greater than the 0.05 when, in fact, the null hypothesis is true, thus increasing the probability that one test will be significant and the other not.

Examples of the Error

Many examples of this mistake exist in practice (reviewed generally in [88,89] and in some letters to editors about childhood obesity specifically[90,91]). Two examples specific to childhood obesity are below.

Researchers investigated a health promotion model for children.[92] The results showed that BMI standard deviations scores (BMI SDS) decreased significantly in the health promotion group ($p<0.001$), but did not differ significantly in the control group. However, the median change in both groups was −0.1 BMI SDS units, for a between-group difference in medians of 0.[93]

In another study, researchers compared the effectiveness of family-based interventions for childhood obesity, in which one intervention included parents, the other included both parents and children, and the control was follow-up only.[94] Although the researchers conducted the appropriate among-group tests that were not statistically significant, the authors nonetheless made conclusions based on the within-group significance of the ‘parents and children’ group.[95]

Recommendations

Authors who compare an outcome measurement with baseline should always be clear that this does not tell them anything about differences between groups for an outcome measure, and does not provide reliable evidence of the effect of the intervention (see “Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time”). The between group comparisons in the case of randomized interventions can be tested several ways, including incorporating the baseline measurement as a covariate, conducting a repeated measures ANOVA, or using follow-up only measurements in the case of randomization (though this would be underpowered compare to including the baseline measurement), among others.

Inferential Error: Equating ‘No Statistically Significant Difference’ with ‘Equally Effective’

Error Description

When an active comparator, instead of a placebo, is used to test a novel intervention’s effectiveness on obesity and there is a null result, conclude that the interventions had ‘equal effectiveness’ rather than ‘were not statistically significantly different.’

Explanation of the Error and Why the Practice is Wrong

The use of placebo or no-attention controls can be controversial, especially when an assumed effective intervention exists. On the one hand, the use of a placebo benchmark for new interventions represents a lower, easier-to-beat efficacy standard than comparing to the existing intervention. On the other hand, because of publication biases[96] and other forces that distort the evidence in the published literature[97–103] it cannot be taken for granted that the existing intervention is actually effective, or effective in all populations (c.f. [103] and [104] for discussions about placebo controls). For the present discussion, we simply acknowledge that there are principled reasons why a researcher might want to conduct a placebo-less, head-to-head comparison between two interventions, each of which may be conjectured to have some efficacy.

The claims made from such a design, however, are more nuanced. Consider a situation in which two interventions are being compared and the outcome is weight loss. Here, the usual null hypothesis is that the two interventions have the same effectiveness and thus the average weight loss is the same across groups. The complementary alternative hypothesis is that the novel intervention produces either superior or inferior weight loss compared to the existing intervention. This is the setup for a *superiority trial*.[105] In practice, when the null is rejected, the question of superiority or inferiority is easily settled by the direction of the observed effect; however, the null will only be rejected in sufficiently powered research with either large sample sizes when effect sizes are small, or when there are large effect sizes. On the other hand, if the study has low power and small true effects, one can almost *a priori* guarantee a non-significant result. When there is no statistically significant difference between groups, and particularly in situations where both groups improved from baseline, researchers may make two mistakes. First, authors may conclude that the change from baseline is evidence that the intervention worked at all; however, without the appropriate placebo control it is always possible that the improvement was coincidental or a statistical artifact like RTM (see “Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time”). Second, because the two groups were not significantly different, authors may incorrectly ‘accept the null’ when discussing non-significant differences between groups and declare ‘equal effectiveness’ between a novel intervention and the existing intervention, when in fact ‘unequal effectiveness’ is also compatible with the data (Figure, Cases 2–4).

Examples of the Error

In a randomized comparison of therapist-led (TLG) and self-help groups (SHG), “[n]o significant between-group differences were detected in the children’s changes in adiposity or dietary intake after 6 and 24 months”; but this does not necessarily mean that “the TLG and SHG intervention groups appear to be equally effective in improving long-term adiposity and dietary intake in obese children.”[106] Similarly, if “[c]hild BMIz outcomes were not statistically different between the two groups (F = 0.023, p = .881)” then one should not necessarily claim that “[b]oth telemedicine and structured physician visit[s] may be feasible and acceptable methods of delivering pediatric obesity intervention to rural children.”[107] Even with a highly significant “reduction in the ZBMI in both groups (P<0.0001), without [a] significant difference between them (P=0.87)” one should not claim that “fixed diet plan[s] and calorie-counting diet[s] led to a similar reduction of ZBMI”[108] because there is no non-treatment or placebo comparator.

Recommendations

If a researcher wants to show that a novel intervention is superior to an existing intervention and furthermore that it is effective in its own right, the way to do this is to conduct a three-arm trial comparing the novel intervention, the existing intervention, and a placebo or non-treatment control. If the two interventions are indeed effective, demonstrating effectiveness versus placebo should not be difficult. However, if both interventions are effective, and the difference in effectiveness between two interventions is small, very large sample sizes may be necessary to detect a difference, which could make the study impractical.

A researcher might *a priori* decide to investigate whether the novel intervention is ‘equally effective’ or ‘not worse’ than the existing intervention. For either goal, a superiority trial should not be used. Rather, the trial must be set up as an *equivalence trial* or a *non-inferiority trial*, respectively.[109] Non-inferiority trials use a different, one-sided null, and as a result a rejected null would be interpreted as “the novel intervention is no worse than Δ% less effective than the existing intervention”, where Δ is small and determined *a priori*. An equivalence trial is similar, but two-sided: “the novel intervention is no better or worse than Δ% effective than the existing intervention” (Figure, Case 1). However, because of this design choice, a non-inferiority trial cannot be used to show superiority over an existing intervention.[110] An extension of the CONSORT guidelines is available for reporting non-inferiority and equivalence trials.[111]

As always, the question to be answered should be determined before the research begins and the corresponding proper design must be implemented. Trying to utilize a superiority trial as a non-inferiority or equivalence trial or vice-versa is unacceptable. Results that are compatible with “equally effective” are also compatible with “equally ineffective.”

Inferential Error: Ignoring Intervention Study Results in Favor of Observational Analyses

Error Description

If the intervention does not produce better results than the control, ignore or underemphasize the original intervention design in favor of observational correlations of intervention-related factors with outcomes.

Explanation of the Error and Why the Practice is Wrong

When differences between the intervention and control groups are not detected, researchers may choose to ignore the original design and instead test for and emphasize associations to support their causal claims. For instance, the control group may be ignored, and regressions between intervention compliance (e.g., number of intervention sessions attended) and outcomes might only be tested within the intervention group. Or, the groups may be pooled, and some aspect of the treatment (e.g., number of fruit and vegetable servings) might be tested for its relation to outcomes across all participants. This vitiates the more sound, between-group inferences and removes intervention assignment, thereby undermining causal inference and forfeiting the strengths of a randomized trial. This becomes even more concerning when comparison groups are formed using characteristics that are measured post-randomization.[112] The dropping or pooling of comparator groups to focus on changes over time can be problematic regardless of whether the interventions were randomized (e.g., a randomized trial) or not (e.g., a quasi-experiment), and is therefore related to Errors “Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time” and “Basing Conclusions on Tests for Significant Differences from Baseline”. Secondary or exploratory analyses can lead to important new hypotheses, but selectively ignoring data (e.g., the control group) or study design (e.g., randomization) limits causal inference of the study *as designed,*[113] and may be misleading if the primary, between-group design is ignored or underemphasized.

Examples of the Error

The Healthy Schools Program (HSP) is a national program that provides schools with tools to design healthy food and physical activity environments. To examine the effectiveness of the program for reducing the prevalence of childhood overweight and obesity, a study was conducted comparing schools with the HSP intervention and propensity-score matched controls.[114] Although the study found no differences between the two groups on the prevalence of overweight and obesity, the authors claimed “clear” effectiveness of the HSP based on secondary analyses of the participating schools (excluding the controls), which demonstrated a mild dose-response relationship between years of contact with the program and reduction in prevalence of overweight and obesity. The investigators deemed the intervention as “evidence based” and concluded that it was, “an important means of supporting schools in reducing obesity” despite the lack of evidence from the between-group comparison. A dose response of the intervention is one potential explanation for the within-group results, but, given the non-significant between-groups analysis, a compelling alternative explanation for the association is that the schools that accepted more of the intervention were different from those that accepted less.

Another example investigated the effect of once or twice per week delivery of a family-based intervention.[115] Although no differences were seen between the two versions of the program, the authors concluded that "higher attendance, as a proportion of available sessions, leads to better outcomes for children." This conclusion was based on pooling the two groups and looking for associations among proportion of attendance and outcomes. As in the previous example, it is possible that there is an inherent difference between children who adhere and those who do not. Indeed, in this case, equal adherence to a proportion of sessions meant that the twice-per-week group had to attend twice as many sessions as the once-per-week group, and yet twice the exposure (as randomized) did not result in a difference between groups.

Recommendations

Rigorously conducted and adequately powered studies with non-significant between-group results still provide useful information about the effectiveness – or lack thereof – of the interventions. Ignoring the primary results in favor of testing associations within subgroups or using post-randomization tests is discouraged. These exploratory analyses can be integral to investigating what characteristics of children or the interventions might lead to effectiveness, but the analyses need to be communicated clearly, with appropriate limitations cited, and making it clear to the reader that conclusions are from associations and do not have the strength of trial results.

Inferential Error: Using One-sided Testing for Statistical Significance

Error Description

If statistically significant results are not achieved with a two-sided test at the conventional 0.05 significance level, but the p-value is less than 0.10 and the effect estimate is in the preferred direction, switch to a one-sided test and it will be significant.

Explanation of the Error and Why the Practice is Wrong

Let us take a scenario in which a researcher uses a two-sided t-test at the 5% significance level (α=0.05) to assess the between-group difference in BMI as the primary outcome of a childhood obesity intervention. The researcher expects that the intervention group will have a lower post-intervention mean BMI than the control group, with a formal null hypothesis that the intervention group is equal to the control group. Contrary to the investigator's hopes, the two-sided p-value turns out to be 0.08 in the favored direction, thus failing to reject the null hypothesis. However, because the researcher is confident that the effect can only be in one direction, the initial analysis plan is abandoned (see "Following the Forking Paths") in favor of a one-sided test. The null hypothesis for this new test is now that the intervention is worse than or equal to the control, while the alternative hypothesis is that the intervention is better than the control. The one-sided test no longer guards against a mistaken null hypothesis rejection in the opposite direction, so practically speaking for this case the obtained p-value is cut in half when the difference is in the favored direction. The p-value is now 0.04: statistically significant.

When researchers are not formally testing non-inferiority (see "Equating 'No Statistically Significant Difference' with 'Equally Effective'"), the described approach is wrong for at least two reasons.[1] First, unless one is explicitly utilizing Bayesian statistics with subjective priors (not discussed herein), results should be independent of the researcher's expectations. The results require "a respect that transcends the stories they can tell about how they came to do the experiment, which they call 'theories.'"[1,116] Although a researcher is not interested in one of the two directions, future readers may come up with another theory that hypothesizes the opposite effect or no effect at all, and reporting and interpreting results in only one direction limits the utility of the results for future scrutiny. Second, the research may result in a large difference in the unexpected direction, yet one-sided tests do not differentiate between no effect and large effects in the undesired direction. Researchers using a one-sided test may then be tempted to offer an explanation for the large effect in the unexpected direction, which violates the assumptions of the one-sided test. One-sided tests only test a single direction, and any attempt to interpret the effect in the unexpected direction essentially has a type I error rate of 10% (5% in each direction) instead of the stated 5%.

Examples of the Error

In some cases, authors justified the use of one-sided tests by stating that their hypotheses are directional to begin with.[117] Yin et al.[118] specifically argued that their prior study results justified testing new results only in the direction consistent with their prior results. Others reported one-sided tests only for some outcomes.[119] Based on the manner in which statistics were reported, it seems likely that one-sided tests utilized in some childhood obesity interventions remain partly disclosed[36] or undisclosed[120] because the authors did not state whether one- or two-sided tests were implemented. For partial disclosure, Siegel et al.[36] reported one-sided tests for some analyses, but did not specify for others. In one ambiguous example, change in BMIz was reported with a confidence interval of (−0.09, 0.02) that contained the null value (Figure, Cases 2–4), but also reported a statistically significant p-value, which is impossible if the confidence interval was constructed from the same statistical procedures. However, statistical significance was possible for that example with a one-sided test. Detecting non-disclosure is more difficult. Kilanowski & Gordon[120] analyzed differences in changes in body weight and BMI between intervention and comparison groups and reported Rank Sum z-values that would provide p-values of 0.107 and 0.121 in two-sided tests, but the authors reported p-values of 0.05 and 0.059 –half of the two-sided (within rounding error), which is consistent with an undisclosed one-sided test.

What is recommended

Long-standing literature on this issue[1,121] emphasizes that a one-sided test in an RCT is not reasonable, except for a non-inferiority trial (see "Equating 'No Statistically Significant Difference' with 'Equally Effective'"). Apart from non-inferiority trials, regardless of justifications, one-sided tests do not seem defensible choices. In all cases, the decision of which tests to use should be stated *a priori* to guard against post hoc decision-making (see "Following the Forking Paths").

Inferential Error: Stating that Effects are Clinically Significant Even Though They Are Not Statistically Significant

Error Description

When results are not statistically significant, ignore the statistical tests in favor of making optimistic conclusions about whether the effects are clinically significant (or represent a ‘real-world difference,’ have ‘public health relevance,’ or would create a ‘meaningful impact’).

Explanation of the Error and Why the Practice is Wrong

> “Clinical significance may have to be adjudicated by collective groups. This remains in the eye of the beholder, but as a minimum there is no clinical significance without statistical significance.”[122]

With so much time, energy, and personal commitment invested in an intervention, it may be hard to accept that an intervention was not as unambiguously effective as hoped. This is especially true when statistically non-significant results have a large mean difference, confidence intervals that include impressively large effects, or a p-value close to the threshold of significance, making the results still seem ‘promising.’ The inferential error of ignoring statistical significance in favor of this optimism may reflect at least two misunderstandings of statistical tests.

‘Statistical significance’ here refers to the use of null hypothesis testing as the basis for statistical inference, in which the null hypothesis assumes no difference between groups. There is much discussion about whether[123] and how to use null hypothesis significance testing,[123,124] including whether 0.05 is the appropriate cutoff for statistical significance. Herein, we do not debate those issues, but address studies that use null hypothesis significance testing, of which there are many. However, the error described here can be generalized to the practice of ignoring whatever inferential procedures the researchers have initially chosen.

A common misunderstanding is that failing to reject the null hypothesis (often, when p>0.05) means that we conclude that there is no difference – a fallacy known as ‘accepting the null’ (see “Equating ‘No Statistically Significant Difference’ with ‘Equally Effective’”). Rarely are studies conducted in which we try to conclude that there is no difference, which may look like Case 1 in the Figure. Instead, statistically non-significant results could indicate there genuinely is no or minimal effect (i.e., the null is true), or that there is an effect that investigators were unable to observe in the present study. Authors must conclude there is insufficient evidence to reject that the two groups are the same, but instead authors sometimes inappropriately declare such results as ‘clinically meaningful,’ despite failing to meet the pre-specified threshold to conclude the groups are different at all.

A second misunderstanding is of summary statistics. Notably, researchers committing this error often refer to the point estimate (such as the sample mean) to declare clinical significance. We can use confidence intervals – which are directly related to p-values – to illustrate the problem with this logic. Confidence intervals are constructed in a way that a certain percentage (e.g., 95%) of intervals calculated the same way would contain the true effect value under some assumptions. If we take an example where the null hypothesis is 'zero difference between groups', then if the interval does not include zero we reject the null hypothesis, which is also consistent with $p<0.05$ (Figure, cases 5–7). However, if the interval does include zero, the fact that more of the interval is to one side of zero should not be used as evidence to support rejecting the null hypothesis in this statistical framework (Figure, Cases 2–4). Touting the mean difference (Case 4) or upper confidence limit (Case 3) as 'clinically meaningful' despite having a null or deleterious lower confidence limit, confuses that we have limited information about the magnitude of the effect (i.e., the effect *could* be clinically meaningful) with information that the effect is *likely* to be clinically meaningful, despite the effect potentially being clinically insignificant or even deleterious.

As the introductory quotation for this error makes clear, defining clinical significance is a subjective exercise, just as is defining thresholds for statistical significance. A common convention with statistical significance is $p<0.05$; but for clinical relevance, it is often unclear just how much an outcome has to change before the effects become meaningful. In public health, a minuscule difference may be declared important when integrated over an entire population; for individual health, results might have to be much more striking before affecting clinical practice. Regardless, for any given application, the threshold should be established *a priori*. If establishing clinical significance is the goal then researchers have an alternative hypothesis of interest other than just 'not null.' This concept is illustrated by the 'clinical significance' region in the Figure. Only Case 7 is clearly consistent with rejecting values below clinical significance, and is also statistically significant. For Case 6, we cannot reject values in the clinically non-significant range despite being statistically significantly different from the null with a point estimate above clinical significance.

A corollary is that we must not ignore the clinical triviality of some statistically significant results, such as when the entire 95% confidence interval is below the threshold of clinical significance. That is, we cannot assume clinical significance just because there is statistical significance. Case 5 shows an example where results are statistically significant, and yet fail to include clinical significance in the confidence interval.

We note that comparing confidence intervals to clinical thresholds is related to an approach called magnitude-based inference[125,126] popularized in the field of sports science. It has seen its fair-share of debate on whether it should be implemented[127–130]. Therefore, we encourage readers to use caution with that approach.

Examples of the Error

Ignoring statistical tests in favor of clinical significance manifests in several different ways. Sometimes these reports acknowledge that the intervention did not have a statistically significant effect on the primary body composition outcome, but contend that the effect size was none-the-less clinically significant.[131,132] Non-significant interventions have been said to bring "effective results for the prevention of childhood obesity,"[133] to be "a promising … strategy for preventing childhood obesity,"[134] or "can improve … key weight related behaviors."[135] Other investigators also recognized the lack of statistical significance at the primary experimental design level, but pointed out that a change in the desired direction was significant in potentially non-pre-specified subgroups (i.e., Errors “Changing the Goal Posts” and “Following the Forking Paths),[136,137] or significant among those who received more exposure to the intervention (i.e., Error “Ignoring Intervention Study Results in Favor of Observational Analyses”).[138]

Recommendations

Defining success in advance is important to prevent this error. Researchers should be discouraged from using ‘clinical significance’ to circumvent statistical significance. Clinical significance should be defined *a priori*, and built into power analyses and the statistical analysis plan, and success only declared if non-clinically meaningful values are rejected in appropriate statistical tests. If researchers analyze results without using the common approach of statistical significance thresholds (e.g., by using Bayesian analysis instead), it is still important to state the analysis plan and criteria for success *a priori*. If traditional statistical significance (e.g., evidence the effect is non-zero) is the goal of the research, then the goal of statistical significance should still be defined *a priori*. These recommendations are facilitated by study registration, which is increasingly becoming required.[85]

Discussion and Conclusions

> “[I]n science, three things matter: the data, the methods used to collect the data (which give them their probative value), and the logic connecting the data and methods to conclusions. Everything else is a distraction.”[139]

Reducing childhood obesity is of undeniable importance. So, too, is the need for greater rigor, reproducibility, and transparency in the implementation of much scientific research.[8,139] Our aim here is to be constructive and help the research community interested in this goal to better evaluate, generate, and describe the evidence on strategies to treat or prevent obesity, with an emphasis on childhood obesity interventions. We also hope that this list will lead to elevated – yet healthy – skepticism about claims of effectiveness of childhood obesity interventions. Doubt and skepticism expressed in good faith should be seen as important to advancing science and finding real solutions.[140] White Hat Bias (“bias leading to the distortion of information in the service of what may be perceived to be righteous ends”[97]) risks diverting attention from the important goal, in this case decreasing childhood obesity. Indeed, researchers more readily overlook practices that undermine the validity of research when paired with a justifiable motive,[141] reinforcing the importance of focusing on the rigor of the science itself apart from the perceived importance of the topic. Although we have focused on these errors in the childhood obesity intervention literature, we recognize that these same errors can and do occur in obesity intervention studies in general[7] and in domains other than obesity. As such, this paper may also be useful beyond the focus of childhood obesity.

We make here several recommendations on how to avoid the errors, with full transparency that our recommendations are face-valid, are not necessarily newly proposed by us, and may not yet have been formally proven to improve the practice of science. Some of the errors described herein may be prevented by better statistical and design education, but may also be prevented by substantial inclusion of individuals formally trained in statistics and design as part of an interdisciplinary team. Pre-registration of studies, such as with ClinicalTrials.gov or the Open Science Framework can help researchers plan *a priori* how they will be conducting and analyzing a study, which decouples data-analysis decisions from data-collection decisions, and gives the authors a predefined roadmap to follow for their primary outcomes. However, in at least one case, having both statistical expertise and pre-registration was not sufficient to avoid some errors (c.f. [61] about [60]).

Some more explicit techniques that separate the methods and analysis from the conclusions have been proposed, including: 1) registered reports, in which authors pre-register their design and analysis, and acceptance is dependent on adherence to or justifying deviation from the pre-registered plan. It is important to note here the idea of justified deviation. In one example, the authors report they mistakenly included BMI percentile as opposed to BMIz in their registration, and clarified the distinction well before the final analysis, and still reported both outcomes to remain true to the registration.[142] Journals that require pre-registration implement an informal version of registered reports, but the checking of registrations against the final publications has not been as robust as it should be for this approach to be effective in general,[47] with sharing of protocols in addition to registration resulting in more clarity in selective outcome reporting.[143] 2) Separate peer review of methods and conclusions, in which the methods of a study are reviewed prior to seeing results or conclusions, so acceptance decisions are first dependent on the methodology, which give data their meaning. 3) Triple blinded studies, in which the subjects, the evaluators, and the statisticians are blinded. Such blinding can be particularly difficult in obesity interventions, but ethically masking interventions and comparators, the interventionists, the evaluators, and the data analysts as much as possible can better separate expectations from conclusions. And, 4) completely separating the intervention, evaluation teams, and data analysis teams: an extension of our last point. The services of an independent data management and analysis coordinating center may be particularly useful to control inferential errors such as “Changing the Goal Posts” and “Following the Forking Paths”, which are difficult for the reviewer and other readers to detect from the published paper alone. The passion that researchers need to have to overcome the regulatory, community, and interpersonal hurdles of working with children risks biasing the intervention and analysis because we researchers are human and, despite our best efforts, our expectations and desires may influence the research. Putting up firewalls between the components of an intervention may decrease the influence of these expectations and desires.

Finally, as researchers, our commitment should first be to the truth. Authors, reviewers, editors and readers all can play a role in assuring that fidelity is maintained in conducting research and conveying research findings. We hope that our paper may help to recognize flaws that occur in research on interventions aimed at reducing childhood obesity. It may serve as a checklist to complement existing guidelines (e.g., [80]) and compendia of errors and biases (e.g., [144]) in the spirit of literature showing that simple checklists can be helpful in reducing error rates.[145] It is vital to ensure invalid methodology and interpretations are avoided so that we can identify and support the most promising childhood obesity interventions, while avoiding those that are clearly ineffective.

Acknowledgements

Funded in part by NIH grants R25DK099080, R25HL124208, P30DK072476, and P30DK040561; USDA/ARS under Cooperative Agreement No. 58-3092-5-001; and DARPA grant D17AC00001. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of these agencies or any other organization.

References

1. Burke CJ. Further Remarks on One-Tailed Tests. Psychol Bull. 1954;51(6):587–590.
2. Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM, Armstrong SC. Prevalence of Obesity and Severe Obesity in US Children, 1999–2016. Pediatrics. 2018.
3. Arteaga SS, Esposito L, Osganian SK, Pratt CA, Reedy J, Young-Hyman D. Childhood obesity research at the NIH: Efforts, gaps, and opportunities. Translational Behavioral Medicine. 2018;8(6):962–967. [PubMed: 30329138]
4. Wood AC, Wren JD, Allison DB. The Need for Greater Rigor in Pediatric Obesity Research. JAMA Pediatrics. In press
5. Bauchner H. Notice of retraction: Wansink B, Cheney MM. Super bowls: serving bowl size and food consumption. JAMA. 2005;293(14):1727–1728. JAMA. 2018;320(16):1648–1648. [PubMed: 15827310]
6. Hart A. Common statistical mistakes. Maternal & Child Nutrition. 2012;8(4):421–422. [PubMed: 22937825]
7. George BJ, Beasley TM, Brown AW, et al. Common scientific and statistical errors in obesity research. Obesity (Silver Spring). 2016;24(4):781–790. [PubMed: 27028280]
8. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: A tragedy of errors. Nature. 2016;530(7588):27–29. [PubMed: 26842041]
9. Stevens J, Taber DR, Murray DM, Ward DS. Advances and Controversies in the Design of Obesity Prevention Trials. Obesity. 2007;15(9):2163–2170. [PubMed: 17890483]
10. Hebert JR, Ma Y, Clemow L, et al. Gender differences in social desirability and social approval bias in dietary self-report. Am J Epidemiol. 1997;146(12):1046–1055. [PubMed: 9420529]
11. Havermans N, Vanassche S, Matthijs K. Methodological Challenges of Including Children in Family Research: Measurement Equivalence, Selection Bias and Social Desirability. Child Indic Res. 2015;8(4):975–997.
12. Natarajan L, Pu M, Fan J, et al. Measurement error of dietary self-report in intervention trials. Am J Epidemiol. 2010;172(7):819–827. [PubMed: 20720101]
13. Harnack L, Himes JH, Anliker J, et al. Intervention-related bias in reporting of food intake by fifth-grade children participating in an obesity prevention study. Am J Epidemiol. 2004;160(11):1117–1121. [PubMed: 15561991]
14. Taber DR, Stevens J, Murray DM, et al. The effect of a physical activity intervention on bias in self-reported activity. Ann Epidemiol. 2009;19(5):316–322. [PubMed: 19230711]
15. Paez KA, Griffey SJ, Thompson J, Gillman MW. Validation of self-reported weights and heights in the avoiding diabetes after pregnancy trial (ADAPT). BMC Med Res Methodol. 2014;14:65. [PubMed: 24886128]
16. Harrington KF, Kohler CL, McClure LA, Franklin FA. Fourth graders' reports of fruit and vegetable intake at school lunch: does treatment assignment affect accuracy? Journal of the American Dietetic Association, 109(1), 36–44. 2009. [PubMed: 19103321]
17. Pronk NP, Crain AL, VanWormer JJ, Martinson BC, Boucher JL, Cosentino DL. The use of telehealth technology in assessing the accuracy of self-reported weight and the impact of a daily: immediate-feedback intervention among obese employees. International journal of telemedicine and applications. 2011;2011:4.
18. Caballero B, Clay T, Davis SM, et al. Pathways: a school-based, randomized controlled trial for the prevention of obesity in American Indian schoolchildren. Am J Clin Nutr. 2003;78(5):1030–1038. [PubMed: 14594792]

19. Klesges RC, Obarzanek E, Kumanyika S, et al. The Memphis Girls' health Enrichment Multi-site Studies (GEMS): an evaluation of the efficacy of a 2-year obesity prevention program in African American girls. Arch Pediatr Adolesc Med. 2010;164(11):1007–1014. [PubMed: 21041593]

20. Wiltheiss GA, Lovelady CA, West DG, Brouwer RJN, Krause KM, Ostbye T. Diet Quality and Weight Change among Overweight and Obese Postpartum Women Enrolled in a Behavioral Intervention Program. J Acad Nutr Diet. 2013;113(1):54–62. [PubMed: 23146549]

21. Omorou AY, Langlois J, Lecomte E, Vuillemin A, Briançon S, Group PT. Adolescents' Physical Activity and Sedentary Behavior: A Pathway in Reducing Overweight and Obesity: The PRALIMAP 2-Year Cluster Randomized Controlled Trial. Journal of Physical Activity and Health. 2015;12(5):628–635. [PubMed: 25393601]

22. Arija V, Villalobos F, Pedret R, et al. Effectiveness of a physical activity program on cardiovascular disease risk in adult primary health-care users: the "Pas-a-Pas" community intervention trial. BMC Public Health. 2017;17(1):576. [PubMed: 28619115]

23. Aittasalo M, Jussila A-M, Tokola K, Sievänen H, Vähä-Ypyä H, Vasankari T. Kids Out; evaluation of a brief multimodal cluster randomized intervention integrated in health education lessons to increase physical activity and reduce sedentary behavior among eighth graders. BMC Public Health. 2019;19(1):415. [PubMed: 30995905]

24. Dhurandhar NV, Schoeller D, Brown AW, et al. Energy balance measurement: when something is not better than nothing. International Journal of Obesity. 2015;39(7):1109–1113. [PubMed: 25394308]

25. Schoeller DA, Thomas D, Archer E, et al. Self-report-based estimates of energy intake offer an inadequate basis for scientific conclusions. Am J Clin Nutr. 2013;97(6):1413–1415. [PubMed: 23689494]

26. Klesges LM, Baranowski T, Beech B, et al. Social desirability bias in self-reported dietary, physical activity and weight concerns measures in 8- to 10-year-old African-American girls: results from the Girls Health Enrichment Multisite Studies (GEMS). Preventive medicine. 2004;38 Suppl:S78–87. [PubMed: 15072862]

27. Galton F. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland. 1886;15:246–263.

28. Bland JM, Altman DG. Statistic Notes: Regression towards the mean. Bmj. 1994;308(6942):1499–1499. [PubMed: 8019287]

29. Bland JM, Altman DG. Statistics Notes: Some examples of regression towards the mean. Bmj. 1994;309(6957):780–780. [PubMed: 7950567]

30. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. BMJ. 2000;320(7244):1240. [PubMed: 10797032]

31. Burke RM, Meyer A, Kay C, Allensworth D, Gazmararian JA. A holistic school-based intervention for improving health-related knowledge, body composition, and fitness in elementary school students: an evaluation of the HealthMPowers program. Int J Behav Nutr Phys Act. 2014;11:78. [PubMed: 24969618]

32. 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 Phy. 2015;12:56.

33. Yeh Y, Hartlieb KB, Danford C, Catherine Jen KL. Effectiveness of Nutrition Intervention in a Selected Group of Overweight and Obese African-American Preschoolers. J Racial Ethn Health Disparities. 2018;5(3):553–561. [PubMed: 28699045]

34. Cockrell Skinner A, Goldsby TU, Allison DB. Regression to the Mean: A Commonly Overlooked and Misunderstood Factor Leading to Unjustified Conclusions in Pediatric Obesity Research. Child Obes. 2016;12(2):155–158. [PubMed: 26974388]

35. Yeh Y, Hartlieb KB, Danford C, Jen KC. Correction to: Effectiveness of Nutrition Intervention in a Selected Group of Overweight and Obese African-American Preschoolers. J Racial Ethn Health Disparities. 2018;5(3):562. [PubMed: 29076062]

36. Siegel RM, Pitner HE, Kist C, et al. Obese children in a community YMCA "Fun 2B Fit" program have a reduction in BMI Z-scores. Clin Pediatr (Phila). 2014;53(7):698–700. [PubMed: 24137026]

37. Allison DB. Comment on “School-based health center-based treatment for obese adolescents: feasibility and body mass index effects.”. 2018; https://hypothes.is/search?q=tag%3APubMedCommonsArchive+25259781. Accessed 2018 OCT 24.

38. Hannon BA, Thomas DM, Siu C, Allison DB. The claim that effectiveness has been demonstrated in the Parenting, Eating and Activity for Child Health (PEACH) childhood obesity intervention is unsubstantiated by the data. Br J Nutr. 2018;120(8):958–959. [PubMed: 30160224]

39. Streiner DL. Statistics Commentary Series: Commentary #16-Regression Toward the Mean. J Clin Psychopharmacol. 2016;36(5):416–418. [PubMed: 27496345]

40. Levin JR. An Improved Modification of a Regression-toward-the-Mean Demonstration. The American Statistician. 1993;47(1):24–26.

41. Watson PM, Dugdill L, Pickering K, et al. Service evaluation of the GOALS family-based childhood obesity treatment intervention during the first 3 years of implementation. BMJ Open. 2015;5(2):e006519.

42. Stice E, Shaw H, Marti CN. A meta-analytic review of obesity prevention programs for children and adolescents: the skinny on interventions that work. Psychol Bull. 2006;132(5):667–691. [PubMed: 16910747]

43. Chambers DW. Thinking in a straight line. J Am Coll Dent. 2013;80(3):29–40. [PubMed: 24283034]

44. Bagherniya M, Sharma M, Mostafavi Darani F, et al. School-Based Nutrition Education Intervention Using Social Cognitive Theory for Overweight and Obese Iranian Adolescent Girls: A Cluster Randomized Controlled Trial. Int Q Community Health Educ. 2017;38(1):37–45. [PubMed: 29298634]

45. Waters E, Gibbs L, Tadic M, et al. Cluster randomised trial of a school-community child health promotion and obesity prevention intervention: findings from the evaluation of fun ‘n healthy in Moreland! BMC Public Health. 2017;18(1):92. [PubMed: 28774278]

46. 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. 2016;11(6):e16–e17. [PubMed: 25893663]

47. Goldacre B, Drysdale H, Powell-Smith A, et al. The COMPare Trials Project. 2016; www.COMPare-trials.org. Accessed 25 JUL 2018.

48. Lloyd J, Creanor S, Logan S, et al. Effectiveness of the Healthy Lifestyles Programme (HeLP) to prevent obesity in UK primary-school children: a cluster randomised controlled trial. Lancet Child Adolesc Health. 2018;2(1):35–45. [PubMed: 29302609]

49. Barkin SL, Heerman WJ, Sommer EC, et al. Effect of a Behavioral Intervention for Underserved Preschool-Age Children on Change in Body Mass Index: A Randomized Clinical Trial. JAMA. 2018;320(5):450–460. [PubMed: 30088008]

50. Yavchitz A, Boutron I, Bafeta A, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS medicine. 2012;9(9):e1001308.

51. Lee S, Won J, Kim S, Park SJ, Lee H. Spin in Randomised Clinical Trial Reports of Interventions for Obesity. Korean Journal of Acupuncture. 2017;34(4):251–264.

52. Hannon PJ. Experimental social epidemiology: controlled community trials In: Oakes JM, Kaufman JS, eds. Methods in Social Epidemiology. San Francisco: Jossey-Bass/Wiley; 2006:335–364.

53. Campbell MK, Piaggio G, Elbourne DR, Altman DG, Group C. Consort 2010 statement: extension to cluster randomised trials. BMJ. 2012;345:e5661.

54. National Institutes of Health. Group- or Cluster-Randomized Trials (GRTs). Research Methods Resources 2018; https://researchmethodsresources.nih.gov/grt.aspx. Accessed 24 DEC 2018.

55. Heo M, Nair SR, Wylie-Rosett J, et al. Trial Characteristics and Appropriateness of Statistical Methods Applied for Design and Analysis of Randomized School-Based Studies Addressing Weight-Related Issues: A Literature Review. J Obes. 2018;2018:8767315.

56. Brown AW, Li P, Bohan Brown MM, et al. Best (but oft-forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials. Am J Clin Nutr. 2015;102(2):241–248. [PubMed: 26016864]

57. 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;2015:708181.

58. Li P, Brown AW, Oakes JM, Allison DB. Comment on “School-Based Obesity Prevention Intervention in Chilean Children: Effective in Controlling, but not Reducing Obesity”. J Obes. 2015;2015:183528.

59. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: a review of defintions. International Statistics Review. 2009;77(3):378–394.

60. Scherr RE, Linnell JD, Dharmar M, et al. A Multicomponent, School-Based Intervention, the Shaping Healthy Choices Program, Improves Nutrition-Related Outcomes. J Nutr Educ Behav. 2017;49(5):368–379 e361.

61. Wood AC, Brown AW, Li P, et al. A Comment on Scherr et al “A Multicomponent, School-Based Intervention, the Shaping Healthy Choices Program, Improves Nutrition-Related Outcomes”. J Nutr Educ Behav. 2018;50(3):324–325. [PubMed: 29524984]

62. Scherr RE, Linnell JD, Dharmar M, et al. Response to “A Comment on Scherr et al ‘A Multicomponent, School-Based Intervention, the Shaping Healthy Choices Program, Improves Nutrition-Related Outcomes’“. J Nutr Educ Behav. 2018;50(3):326–327. [PubMed: 29524985]

63. Lucan SC. Dramatic Decreases in BMI Percentiles, but Valid Conclusions Can Only Come From Valid Analyses. J Nutr Educ Behav. 2018;50(8):850. [PubMed: 30077581]

64. Corrigendum. J Nutr Educ Behav. 2018;50(8):852. [PubMed: 30077582]

65. Müller I, Schindler C, Adams L, et al. Effect of a Multidimensional Physical Activity Intervention on Body Mass Index, Skinfolds and Fitness in South African Children: Results from a Cluster-Randomised Controlled Trial. International Journal of Environmental Research and Public Health. 2019;16(2):232.

66. Koretz RL. JPEN Journal Club 45. Cluster Randomization. Journal of Parenteral and Enteral Nutrition.0(0).

67. Retraction statement: LA sprouts randomized controlled nutrition, cooking and gardening program reduces obesity and metabolic risk in Latino youth. Obesity (Silver Spring). 2015;23(12):2522. [PubMed: 26524103]

68. Turner EL, Li F, Gallis JA, Prague M, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 1-Design. Am J Public Health. 2017;107(6):907–915. [PubMed: 28426295]

69. Turner EL, Prague M, Gallis JA, Li F, Murray DM. Review of Recent Methodological Developments in Group-Randomized Trials: Part 2-Analysis. Am J Public Health. 2017;107(7):1078–1086. [PubMed: 28520480]

70. Wasserstein RL, Lazar NA. The ASA’s Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129–131.

71. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science. 2011;22(11):1359–1366. [PubMed: 22006061]

72. Simonsohn U, Nelson LD, Simmons JP. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspect Psychol Sci. 2014;9(6):666–681. [PubMed: 26186117]

73. Ioannidis JPA. Commentary: Sequential Discovery, Thinking Versus Dredging, and Shrink or Sink. Epidemiology. 2008;19(5):657–658.

74. Gelman A, Loken E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University. 2013.

75. Gelman A, Loken E. The Statistical Crisis in Science. Am Sci 2014;102(6):460–465.

76. Gadbury GL, Allison DB. Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. PloS One. 2012;7(10):e46363.

77. Rankin J, Ross A, Baker J, O’Brien M, Scheckel C, Vassar M. Selective outcome reporting in obesity clinical trials: a cross-sectional review. Clin Obes. 2017;7(4):245–254. [PubMed: 28557240]

78. ClinicalTrials.gov. Health Promotion in Adolescents in Ecuador (ACTIVITAL). ClinicalTrials.gov 2009; https://clinicaltrials.gov/ct2/show/NCT01004367. Accessed 25 JUL 2018.

79. Andrade S, Lachat C, Cardon G, et al. Two years of school-based intervention program could improve the physical fitness among Ecuadorian adolescents at health risk: subgroups analysis from a cluster-randomized trial. Bmc Pediatr. 2016;16:51. [PubMed: 27102653]

80. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c869. [PubMed: 20332511]

81. van Assen MA, van Aert R, Wicherts JM. Meta-analysis using effect size distributions of only statistically significant studies. Psychological methods. 2015;20(3):293. [PubMed: 25401773]

82. McShane BB, Bockenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. Perspect Psychol Sci. 2016;11(5):730–749. [PubMed: 27694467]

83. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biol. 2015;13(3):e1002106.

84. Kelley GA, Kelley KS. Evidential Value That Exercise Improves BMI z-Score in Overweight and Obese Children and Adolescents. Biomed Res Int. 2015;2015:151985.

85. Hudson KL, Lauer MS, Collins FS. Toward a New Era of Trust and Transparency in Clinical Trials. JAMA. 2016;316(13):1353–1354. [PubMed: 27636028]

86. Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. J Res Educ Eff. 2012;5(2):189–211.

87. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. Trials. 2011;12:264. [PubMed: 22192231]

88. Bland JM, Altman DG. Comparisons within randomised groups can be very misleading. BMJ. 2011;342:d561. [PubMed: 21551184]

89. Bland JM, Altman DG. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. Am J Clin Nutr. 2015;102(5):991–994. [PubMed: 26354536]

90. Allison DB. RE: Statistical Interpretation Error in Metformin Trial Article. Pediatrics. 2017;140(6).

91. McComb B, Frazier-Wood AC, Dawson J, Allison DB. Drawing conclusions from within-group comparisons and selected subsets of data leads to unsubstantiated conclusions: Letter regarding Malakellis et al. Aust N Z J Public Health. 2018;42(2):214. [PubMed: 29281164]

92. Fidanci BE, Akbayrak N, Arslan F. Assessment of a Health Promotion Model on Obese Turkish Children. J Nurs Res. 2017;25(6):436–446. [PubMed: 29099476]

93. Brown AW, Allison DB. Letter to the Editor And response Letter to the Editor and Author Response of Assessment of a Health Promotion Model on Obese Turkish Children. The Journal of Nursing Research, 25(6), 436–446. J Nurs Res. 2018;26(5):373–374.

94. Yackobovitch-Gavan M, Wolf Linhard D, Nagelberg N, et al. Intervention for childhood obesity based on parents only or parents and child compared with follow-up alone. Pediatr Obes. 2018;13(11):647–655. [PubMed: 29345113]

95. Dawson JA, Brown AW, Allison DB. The stated conclusions are contradicted by the data, based on inappropriate statistics, and should be corrected: comment on 'intervention for childhood obesity based on parents only or parents and child compared with follow-up alone'. Pediatr Obes. 2018;13(11):656–657. [PubMed: 30092611]

96. Brown AW, Mehta TS, Allison DB. Publication bias in science: what is it, why is it problematic, and how can it be addressed? In: Jamieson KH, Kahan D, Scheufele DA, eds. The Oxford Handbook of the Science of Science Communication2017:93–101.

97. Cope MB, Allison DB. White hat bias: examples of its presence in obesity research and a call for renewed commitment to faithfulness in research reporting. Int J Obes (Lond). 2010;34(1):84–88; discussion 83. [PubMed: 19949416]

98. Brown AW, Ioannidis JP, Cope MB, Bier DM, Allison DB. Unscientific Beliefs about Scientific Topics in Nutrition–. In: Oxford University Press; 2014.

99. Schoenfeld JD, Ioannidis JP. Is everything we eat associated with cancer? A systematic cookbook review–. The American journal of clinical nutrition. 2012;97(1):127–134. [PubMed: 23193004]

100. Casazza K, Brown A, Astrup A, et al. Weighing the Evidence of Common Beliefs in Obesity Research. Crit Rev Food Sci Nutr. 2015;55(14):2014–2053. [PubMed: 24950157]

101. Casazza K, Fontaine KR, Astrup A, et al. Myths, presumptions, and facts about obesity. New England Journal of Medicine. 2013;368(5):446–454. [PubMed: 23363498]

102. 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(5):1298–1308. [PubMed: 24004890]

103. Stang A, Hense H-W, Jöckel K-H, Turner EH, Tramèr MR. Is it always unethical to use a placebo in a clinical trial? PLoS medicine. 2005;2(3):e72.

104. Boot WR, Simons DJ, Stothart C, Stutts C. The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. Perspectives on Psychological Science. 2013;8(4):445–454. [PubMed: 26173122]

105. Sedgwick P. What is a non-inferiority trial? BMJ: British Medical Journal (Online). 2013;347.

106. Hystad HT, Steinsbekk S, Odegard R, Wichstrom L, Gudbrandsen OA. A randomised study on the effectiveness of therapist-led v. self-help parental intervention for treating childhood obesity. Br J Nutr. 2013;110(6):1143–1150. [PubMed: 23388524]

107. Davis AM, Sampilo M, Gallagher KS, Landrum Y, Malone B. Treating rural pediatric obesity through telemedicine: outcomes from a small randomized controlled trial. J Pediatr Psychol. 2013;38(9):932–943. [PubMed: 23428652]

108. Mendes MD, de Melo ME, Fernandes AE, et al. Effects of two diet techniques and delivery mode on weight loss, metabolic profile and food intake of obese adolescents: a fixed diet plan and a calorie-counting diet. European journal of clinical nutrition. 2017;71(4):549–551. [PubMed: 27650876]

109. Hahn S. Understanding noninferiority trials. Korean J Pediatr. 2012;55(11):403–407. [PubMed: 23227058]

110. Gottlieb S. The FDA should not mandate comparative-effectiveness trials. American Enterprise Institute for Public Policy Research; 2011.

111. Piaggio G, Elbourne DR, Pocock SJ, Evans SJ, Altman DG. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. Jama. 2012;308(24):2594–2604. [PubMed: 23268518]

112. Li P, Brown AW, Dawson JA, et al. Concerning Sichieri R, Cunha DB: Obes Facts 2014;7:221–232. The Assertion that Controlling for Baseline (Pre-Randomization) Covariates in Randomized Controlled Trials Leads to Bias Is False. Obesity Facts. 2015;8(2):127–129. [PubMed: 25871982]

113. Rubin DB. For Objective Causal Inference, Design Trumps Analysis. Ann Appl Stat. 2008;2(3):808–840.

114. Madsen KA, Cotterman C, Crawford P, Stevelos J, Archibald A. Peer Reviewed: Effect of the Healthy Schools Program on Prevalence of Overweight and Obesity in California Schools, 2006–2012. Preventing chronic disease. 2015;12.

115. Khanal S, Welsby D, Lloyd B, Innes-Hughes C, Lukeis S, Rissel C. Effectiveness of a once per week delivery of a family-based childhood obesity intervention: a cluster randomised controlled trial. Pediatric Obesity. 2016;11(6):475–483. [PubMed: 26695932]

116. Cohen J. Some statistical issues in psychological research In: Wolman B, ed. Handbook of clinical psychology. New York: McGraw-Hill; 1965:95–121.

117. Gentile DA, Welk G, Eisenmann JC, et al. Evaluation of a multiple ecological level child obesity prevention program: Switch what you Do, View, and Chew. BMC Med. 2009;7:49. [PubMed: 19765270]

118. Yin ZN, Parra-Medina D, Cordova A, et al. Miranos! Look at Us, We Are Healthy! An Environmental Approach to Early Childhood Obesity Prevention. Childhood Obesity. 2012;8(5):429–439. [PubMed: 23061498]

119. Siwik V, Kutob R, Ritenbaugh C, et al. Intervention in overweight children improves body mass index (BMI) and physical activity. J Am Board Fam Med. 2013;26(2):126–137. [PubMed: 23471926]

120. Kilanowski JF, Gordon NH. Making a Difference in Migrant Summer School: Testing a Healthy Weight Intervention. Public Health Nurs. 2015;32(5):421–429. [PubMed: 25611178]

121. Streiner DL. Statistics Commentary Series: Commentary #12-One--Tailed and Two-Tailed Tests. J Clin Psychopharmacol. 2015;35(6):628–629. [PubMed: 26479225]

122. Krishnan KR. Psychiatric disease in the genomic era: rational approach. Mol Psychiatry. 2005;10(11):978–984. [PubMed: 16077681]

123. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. In: Nature Publishing Group; 2019.

124. Leek J, McShane BB, Gelman A, Colquhoun D, Nuijten MB, Goodman SN. Five ways to fix statistics. Nature. 2017;551(7682):557–559.

125. Batterham AM, Hopkins WG. Making Meaningful Inferences About Magnitudes. International Journal of Sports Physiology and Performance. 2006;1(1):50–57. [PubMed: 19114737]

126. Welsh AH, Knight EJ. “Magnitude-based inference”: a statistical review. Medicine and science in sports and exercise. 2015;47(4):874–884. [PubMed: 25051387]

127. Lakens D. Putting MBI on a formal footing: a comment on The Vindication of Magnitude-Based Inference. Sportscience. 2018;22.

128. Sainani KL. The Problem with “Magnitude-based Inference”. Medicine and science in sports and exercise. 2018;50(10):2166–2176. [PubMed: 29683920]

129. Hopkins WG, Batterham AM. The vindication of Magnitude-Based Inference. Sportscience. 2018;22:19–29.

130. Barker RJ, Schofield MR. Inference about magnitudes of effects. Int J Sports Physiol Perform. 2008;3(4):547–557. [PubMed: 19223677]

131. Bundy A, Engelen L, Wyver S, et al. Sydney Playground Project: A Cluster-Randomized Trial to Increase Physical Activity, Play, and Social Skills. Journal of School Health. 2017;87(10):751–759. [PubMed: 28876473]

132. Greve J, Heinesen E. Evaluating the impact of a school-based health intervention using a randomized field experiment. Economics and Human Biology. 2015;18:41–56. [PubMed: 25898077]

133. Thivel D, Isacco L, Lazaar N, et al. Effect of a 6-month school-based physical activity program on body composition and physical fitness in lean and obese schoolchildren. European Journal of Pediatrics. 2011;170(11):1435–1443. [PubMed: 21475968]

134. Schwartz RP, Hamre R, Dietz WH, et al. Office-based motivational interviewing to prevent childhood obesity: a feasibility study. Arch Pediatr Adolesc Med. 2007;161(5):495–501. [PubMed: 17485627]

135. Smith JJ, Morgan PJ, Plotnikoff RC, et al. Smart-phone obesity prevention trial for adolescent boys in low-income communities: the ATLAS RCT. Pediatrics. 2014;134(3):e723–731. [PubMed: 25157000]

136. Borys JM, Richard P, Ruault du Plessis H, Harper P, Levy E. Tackling Health Inequities and Reducing Obesity Prevalence: The EPODE Community-Based Approach. Annals of Nutrition & Metabolism. 2016;68 Suppl 2:35–38.

137. Woo Baidal JA, Nelson CC, Perkins M, et al. Childhood obesity prevention in the women, infants, and children program: Outcomes of the MA-CORD study. Obesity (Silver Spring). 2017;25(7):1167–1174. [PubMed: 28653498]

138. Hull PC, Buchowski M, Canedo JR, et al. Childhood obesity prevention cluster randomized trial for Hispanic families: outcomes of the healthy families study. Pediatr Obes. 2018;13(11):686–696. [PubMed: 27884047]

139. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: Errors, underlying themes, and potential solutions. Proc Natl Acad Sci U S A. 2018;115(11):2563–2570. [PubMed: 29531079]

140. Allison DB, Pavela G, Oransky I. Reasonable Versus Unreasonable Doubt Although critiques of scientific findings can be used for misleading purposes, skepticism still plays a crucial role in producing robust research. Am Sci. 2018;106(2):84–87.

141. Sacco DF, Brown M, Bruton SV. Grounds for Ambiguity: Justifiable Bases for Engaging in Questionable Research Practices. Sci Eng Ethics. 2018.

142. Paul IM, Savage JS, Anzman-Frasca S, et al. Effect of a Responsive Parenting Educational Intervention on Childhood Weight Outcomes at 3 Years of Age: The INSIGHT Randomized Clinical Trial. JAMA. 2018;320(5):461–468. [PubMed: 30088009]

143. Calmejane L, Dechartres A, Tran VT, Ravaud P. Making protocols available with the article improved evaluation of selective outcome reporting. Journal of clinical epidemiology. 2018;104:95–102. [PubMed: 30196127]

144. Catalogue of Bias Collaboration. Catalogue of Bias. https://catalogofbias.org. Accessed 09 MAY 2019.

145. Gawande A. The Checklist Manifesto. Penguin Books India; 2010.

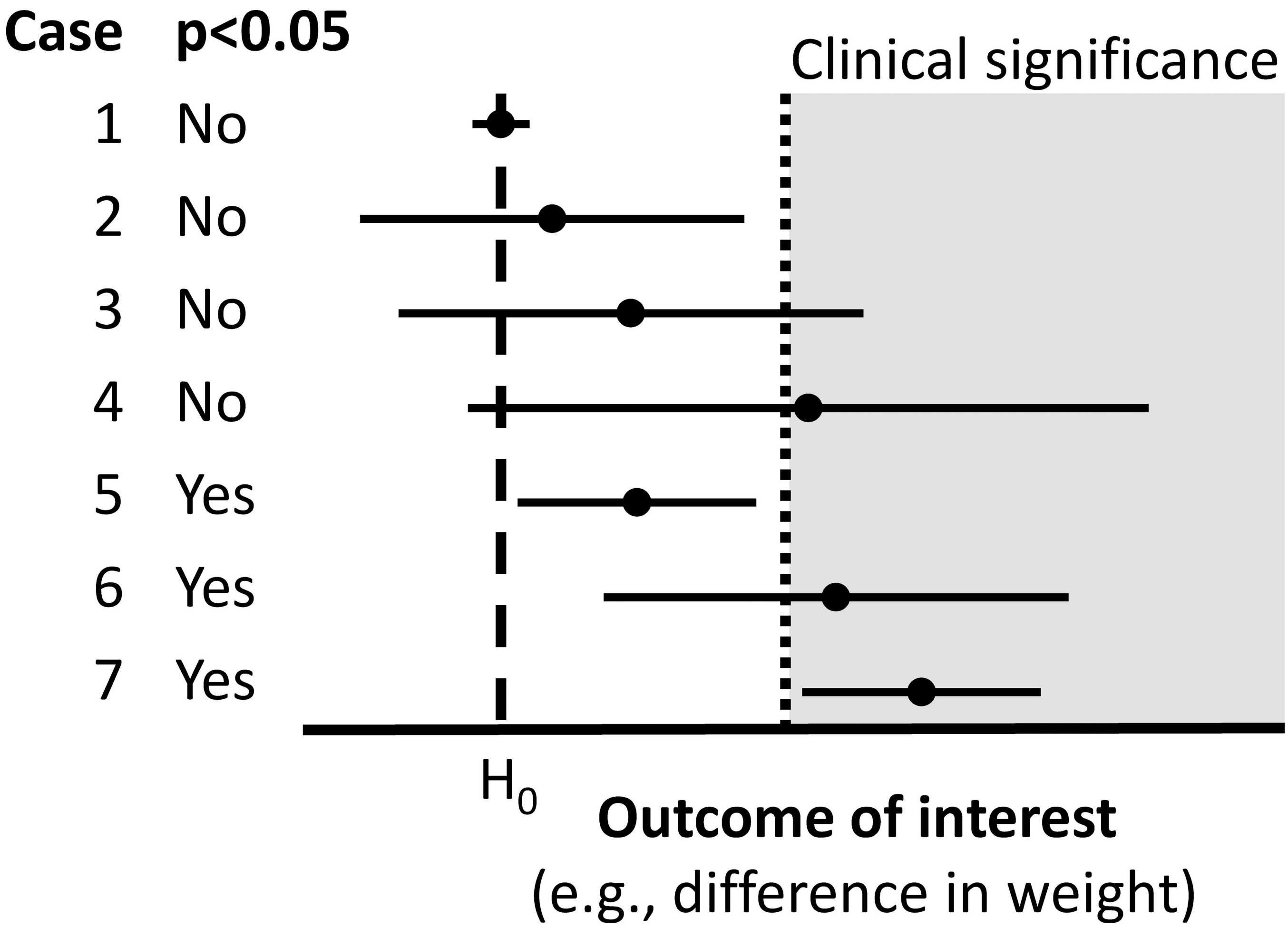

Figure.

Seven hypothetical study results, with point estimates and 95% confidence intervals. H_0 represents the null hypothesis (often representing no differences between groups).

Table:

10 inferential errors, how they may occur, and recommendations for how to avoid them or how to communicate when they are unavoidable.

Inferential error[1]	**Error description**	**Recommendations**[2]
Using Self-Reported Outcomes and Teaching to the Test	Urging the intervention group to change health-related behaviors or conditions, then giving participants a questionnaire that asks about the same health related behaviors and conditions, and ignoring the biases this can induce.	Use objective measurements when possible. If self-report is the only measurement tool available, either forego the measurements entirely, do not emphasize the measurements in the conclusions, or at the very least make the reader aware of the potential for biased results.
Foregoing Control Groups and Risking Regression to the Mean Creating Differences Over Time	Providing an intervention only to individuals preferentially sampled to be either higher or lower than the population mean on some variable – such as children all with high BMI z-scores – and assuming improvements over time are caused by the intervention, rather than a spontaneous tendency for extreme values to revert toward the population average.	Include a control group with the same characteristics as the intervention group. If not available, communicate clearly that subgrouping on extreme values risks the follow-up values being closer to the population average because of regression to the mean rather than an actual effect.
Changing the Goal Posts	Using surrogate or secondary outcomes to make claims of effectiveness for an intervention when a study to test an intervention's effect on obesity yields a non-significant result for the primary outcome.	Focus the report on the pre-registered primary outcome, and communicate intermediate endpoints with great caution.
Ignoring Clustering in Studies that Randomize Groups of Children	Conducting a cluster randomized trial in which groups of children are randomly assigned to experimental conditions, but analyzing the data as though the children were randomized individually.	Always account for clustering in statistical analyses. Have as many clusters as possible, and always more than one cluster per treatment condition.
Following the Forking Paths, Sub-Setting, P-Hacking, and Data Dredging	Trying different analyses with different subsets of the sample or various outcomes and basing conclusions on whatever is statistically significant.	Where appropriate, pre-specify questions and analyses of interest. Be transparent about all analyses conducted, how they were conducted, and whether they were pre-specified. Do not draw definitive conclusions about causal effects from analyses that were not pre-specified or are subsets of many pre-specified analyses uncorrected for multiple testing.
Basing Conclusions on Tests for Significant Differences from Baseline	Separately testing for significant differences from baseline in the intervention and control groups and if the former is significant and the latter is not, declaring the result statistically significant.	Always conduct, report, and emphasize the appropriate between-groups test.
Equating 'No Statistically Significant Difference' with 'Equally Effective'	Concluding that two interventions tested head-to-head had 'equal effectiveness' when there is no statistically significant difference between groups.	Include an appropriate non-intervention control group if absolute effectiveness is of interest. When comparing only two interventions head-to-head, do not presume that changes over time reflect effectiveness. Testing equivalence or non-inferiority between two interventions requires special design and analysis considerations.
Ignoring Intervention Study Results in Favor of Observational Analyses	Drawing conclusions from correlations of intervention-related factors with outcomes, rather than testing the actual intervention against a control as designed.	Report primary, between-group analyses from controlled intervention studies. Clearly communicate that observational findings do not carry the same causal evidence.
Using One-sided Testing for Statistical Significance	Switching to one-sided statistical significance tests to make results statistically significant.	Two-sided tests are typically more appropriate. One-sided tests should not be used. In cases where one insists on their use, the testing approach should be pre-specified and justified.
Stating that Effects are Clinically Significant Even Though They Are Not Statistically Significant	Ignoring the statistical tests in favor of making optimistic conclusions about whether the effects are clinically significant.	Pre-specify what counts as statistically or clinically significant, and be faithful to and transparent about the analysis and interpretation plans. If using statistical significance testing, do not claim that effects have been demonstrated if the effect estimates are not statistically significant, regardless of how large the point estimates are.

[1] The order of errors as presented does not imply a ranking of importance or severity.

[2] In most cases, a common recommendation for hypothesis testing is to preregister or predefine as much as possible. The recommendations below are not meant to discourage hypothesis-generating investigations of the data, but rather to encourage making clear distinctions between hypothesis testing, hypothesis generation, and causal inference.