Published in final edited form as:

Eur J Clin Invest. 2016 December ; 46(12): 985–993. doi:10.1111/eci.12681.

Observational Research Rigor Alone Does Not Justify Causal Inference

Keisuke Ejima^{1,*}, Peng Li^{1,2,*}, Daniel L. Smith Jr.^{3,4,5}, Tim R. Nagy^{3,4,5}, Inga Kadish⁶, Thomas van Groen⁶, John A. Dawson^{1,‡}, Yongbin Yang^{3,4}, Amit Patki², and David B. Allison^{1,2,3,4,5,†}

¹Office of Energetics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

²Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

³Department of Nutrition Sciences, University of Alabama at Birmingham, Birmingham, AL 35294, USA

⁴Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA

⁵Diabetes Research Center, University of Alabama at Birmingham, Birmingham, AL 35294, USA

⁶Department of Cell, Developmental & Integrative Biology, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Abstract

Background—Differing opinions exist on whether associations obtained in observational studies can be reliable indicators of a causal effect if the observational study is sufficiently well controlled and executed.

Materials and methods—To test this, we conducted two animal observational studies that were rigorously controlled and executed beyond what is achieved in studies of humans. In study 1, we randomized 332 genetically identical C57BL/6J mice into three diet groups with differing food energy allotments and recorded individual self-selected daily energy intake and lifespan. In study 2, 60 male mice (CD1) were paired and divided into two groups for a 2-week feeding regimen. We evaluated the association between weight gain and food consumption. Within each pair, one animal was randomly assigned to an S group in which the animals had free access to food. The second paired animal (R group) was provided exactly the same diet that their S partner ate the day before.

Results—In study 1, across all three groups, we found a significant negative effect of energy intake on lifespan. However, we found a positive association between food intake and lifespan among the *ad libitum* feeding group: 29.99 (95% CI: 8.2 to 51.7) days per daily kcal. In study 2,

[†]Correspondence to: Address: 1665 University Blvd, RPHB 140J, Birmingham, AL 35294-0022 Phone: (205) 975-9169 Fax: (205) 975-7536, dallison@uab.edu.

[‡]Current addresses: Department of Nutrition Sciences, Texas Tech University, Lubbock, TX 79409, USA

*These authors contributed equally to this work.

we found a significant ($P=0.003$) group (randomized vs self-selected)-by-food consumption interaction effect on weight gain.

Conclusions—At least in nutrition research, associations derived from observational studies may not be reliable indicators of causal effects, even with the most rigorous study designs achievable.

Keywords

Observational Study; Randomized Controlled Trial; Causality; Research Design; Nutritional Sciences

Introduction

Establishing cause and effect relationships between hypothesized causal factors and outcomes is a key goal in human biomedical and behavioral research. The randomized experimental design, such as the randomized controlled trial (RCT), is the gold standard for establishing such relationships; however, randomized studies are sometimes impractical owing to financial, technical, or ethical concerns. Therefore, the observational study is an essential tool in many fields, particularly in the discovery phase of research.

Regarding observational studies (i.e., studies in which units [subjects] are not randomly assigned to levels of the independent variable under study), agreement exists on the following points:

1. Observational studies have an important place in biomedical and behavioral research.
2. When RCTs are not feasible (not necessarily just not available), the results of observational studies when combined with multiple other sources of evidence may justify a practical conclusion of causation. The classic case of this occurring involves the association of smoking with lung cancer; the classic writing on how one may proceed with such inferences is contained in Sir Austin Bradford Hill's 1965 paper [1].
3. Regardless of how well designed, executed, and analyzed an observational study is, it is always theoretically possible that an association between two variables detected therein is spurious in the sense that the association does not represent a causal effect of one of the variables on the other [2].

In contrast, views among thoughtful scholars differ widely on a practical counterpart to the more theoretical point number 3 above. Stated as a proposition, this point might be phrased as follows:

If an observational study is done extremely well with the variables measured with minimal error, a relatively homogeneous population, great pains taken to minimize specific plausible confounders, and a longitudinal design in which the hypothesized causal variable precedes the hypothesized outcome in time, then for practical purposes, one can rule out the theoretical possibility of a spurious association, and

the observational study alone is sufficient to justify a causal conclusion to a reasonable degree of scientific certainty.

Some authors endorse this view implicitly by their use of causal language to describe findings of association [3–5] and some endorse it explicitly [6], whereas others explicitly eschew it [7]. These widely differing views are expressed clearly in the following two quotations.

It is the position of this task force that rigorous well-designed and well-executed observational studies can provide evidence of causal relationships. (*International Society for Pharmacoeconomics and Outcome Research* [6])

The 12 clinical trials tested 52 observational claims. They all confirmed no claims in the direction of the observational claims...To put it another way, 100% of the observational claims failed to replicate. In fact, five claims (9.6%) are statistically significant in the clinical trials in the opposite direction to the observational claim. (*Young and Karr, after comparing observational and randomized studies of the same hypotheses in mostly nutrition studies* [8])

Ultimately, the reasonableness of the above proposition is an empirical question. Some have tried to address this with meta-analyses [9–12]. Although some of these meta-analyses have shown that observational studies and RCTs tend to obtain similar answers, this is not always the case [13]. Moreover, to our knowledge, almost all the meta-analyses that showed concordance were of treatments applied by physicians to patients with diagnosed disorders. In this case, the populations and the mode of assignment to the hypothesized causal variables may differ enough in the clinical medical setting that the results of these meta-analyses are not necessarily applicable to nonclinical populations, nonclinical settings, or variables such as diet, physical activity patterns, residential locations, career choices, and other factors that individuals can choose without physicians. This may explain why Young and Karr's results quoted above with largely nutritional studies showed an almost complete discordance between the results of the RCTs and the results of the association studies of nutritional factors. Whereas pharmacoepidemiology may benefit from a known biological pathway being targeted under the care of a treating physician with a defined therapy (often singularly assessed), such a scenario is typically absent in nutrition and other lifestyle factor research. Even in the clinical setting, Vandenbroucke distinguished between intended effect, which is expected by the treatment, and unintended effect, which is not expected (such as adverse effect), and he recommended to use observational study only for the causal inference of unintended effect. This claim is reasonable because the population is almost randomly allocated to treatment or control group, which is close to an RCT [14]. Finally, the comparison of separately designed and conducted RCTs with separately designed and conducted observational studies is itself an observational study [15] and is subject to confounding that could either create or mask associations.

Here we report on two unique studies of nutritional factors in which the subjects (mice) were first randomized to be in either an RCT or an observational study of the same nutritional factors and then studied with essentially the same protocol. We chose to do this in mice as a proof of principle assessment, because in mice we can achieve a degree of control for extraneous factors, homogeneity of genotype and living conditions, and measurement

precision that cannot be approached in human studies. Therefore, these mouse studies represent what might be seen as far beyond the plausible upper limit for a rigorous association study in comparison with a rigorous RCT of the same essential question.

Materials and methods

Study 1: Longevity and food consumption in mice

Animals—Over the past few years we have performed a large project (NIA R01AG033682) exploring the effects of weight cycling by repeated bouts of dietary restriction and refeeding on health and longevity outcomes in C57BL/6J mice. Male and female C57BL/6J mice were purchased from the Jackson Laboratory (Bar Harbor, ME, USA) at 6 weeks of age and were acclimated to the facility for 2 weeks. All mice were singly housed for the duration of the study in standard, ventilated mouse cages within a Thoren Rack Mobile Housing System (Thoren Caging Systems, Inc, Hazleton, PA, USA). Animal rooms were maintained at 20–22°C on a 12-hour light-dark cycle from 6:00 AM to 6:00 PM. Animal health was checked daily and moribund animals were euthanized according to the study protocol. Natural death or moribund status termination was recorded to the nearest day. Causes of death were categorized as 1) euthanized for ulcerative dermatitis or similar skin lesions (n=179); 2) euthanized for other reasons, such as disability related to eating or drinking (n=22); 3) found dead (n=125); and 4) died from technical reasons, such as anesthesia (n=3) (Anesthesia was used to perform assessment of visceral adipose tissue using computed tomography in a subset of mice from each treatment group. The first three animals measured at the second timepoint died during the procedure with side effects most likely related to the anesthesia. Therefore we terminated this procedure and made no further measures.). All study protocols were approved by the University of Alabama at Birmingham Institutional Animal Care and Use Committee (IACUC) (#090908909, from 09-18-2009 to the end of the study).

Study design—Beginning at 8 weeks of age, the mice were provided free access to a high-fat diet (45% kcal fat and 20% protein based on D12451 (calorie: 4.73 kcal/g; Research Diets, New Brunswick, NJ) to determine *ad libitum* intake. At 10 months of age, the mice were weighed and the heaviest two-thirds were subsequently randomized by quartile of body weight within each sex into diet groups (see Figure 1, continuing with the high-fat diet feeding until death). The four diet groups were as follows: the ever obese (EO) group, which continued *ad libitum* feeding; the obese weight losers (OWL) group, in which energy intake was restricted by nearly 30% of EO intake (with the vitamin and mineral mix supplemented in the high-fat diet when restriction was >20%, Research Diets #D11022101); the obese weight losers moderate (OWLM) group, in which energy intake was restricted by nearly 20% of EO intake; and the weight cyclers (WC) group, in which energy restriction was enforced by dietary restriction followed by subsequent periods of *ad libitum* refeeding. Energy restriction amounts were adjusted weekly in the OWL and OWLM groups relative to the EO intake and up until approximately 2 years of age, after which the food provisions were maintained due to age-related changes in energy intake in the EO groups as animals approached death. We were able to utilize data from the EO, OWLM, and OWL groups to assess energy intake associations (observational within the EO group) and effects (assignment within the OWLM and OWL groups relative to EO) on the longevity outcomes.

The WC group was excluded from this study because their energy intake varied by time depending on their phase of restriction or refeeding. The OWL diet was designed to reduce the body weight of the mice to a weight comparable to that of mice fed a low-fat diet (10% kcal fat). The OWLM diet was designed to reduce the body weight of the mice to approximately the midpoint of the EO and OWL groups. Owing to the size of the study, the total sample was divided and the study was performed in 2 waves. The two cohorts of animals were separated by approximately 1 calendar year. Experimental wave 1 included both male and female mice, whereas wave 2 focused on males only owing to a higher than expected incidence of ulcerative dermatitis early in life for female mice in experimental wave 1. Additionally, the vitamin A levels in the high-fat diet during experimental wave 2 were reduced to the National Research Council (1995) recommended levels with the same high-fat diet formulation (Research Diets, #D11112301). Following randomization, weekly food intake was measured for animals provided *ad libitum* access (e.g., EO group) along with weekly body weights for all animals. For animals receiving a daily allotment of a restricted amount of food (OWL and OWLM groups), fresh allotments were provided ~1 to 2 hours before lights off, and any food remaining after 24 hours was recorded and discarded before the next day's provisions were given. The current sample size could provide 80% power to detect a small effect size (Cohen's *d* 0.3) regarding lifespan between the OWL and EO groups with a 2-tailed test at an alpha level of 0.05. The experimental protocol is shown in Figure 1. Allocation ratio of mice is approximately 1:2:1 (EO:OWLM:OWL), which was obtained from the sample size calculation conducted for the primary outcome of this experiment.

Outcomes—For the C57BL/6J mice, observed lifespan (days) was recorded as the age when the animals died naturally or were euthanized owing to terminal morbidity. Daily averaged actual energy intake (in calories) after the beginning of randomization (~10 months of age) until death was used to estimate energy intakes among treatment assignments (EO, OWL, and OWLM). We also assessed the following outcomes: general health, body composition, indirect calorimetry (which are not used in this analysis).

Statistical analysis—The characteristics of the mice in different groups were summarized as descriptive statistics (mean and standard deviation) and compared by ANOVA. The lifespans among the three groups were evaluated by using ANOVA and pairwise t-test with Bonferroni correction. A linear regression was performed to examine the association of lifespan and energy intake within each of the three groups.

Study 2: Weight gain and food consumption in young mice

The previous example focused on calorie intake and a long-term, hard endpoint of mortality. However, many human studies are shorter in length and rely on health-related outcomes to predict relative risks. Therefore, we designed and conducted an experiment for this article to emulate such a study. Using data from a food consumption and weight gain experiment, we compared the associations of cause and short-term outcome inferred from the RCT and corresponding observational study.

Animals—Male CD1 mice were purchased from Charles River Laboratories (Portage, MI, USA) at 6 weeks of age and were acclimated to the facility for 2 weeks. All mice were singly housed for the duration of the study in the same condition as described for study 1, and animal care followed IACUC protocols.

Study design—Sixty singly housed, 8-week-old male mice (CD1) were randomly divided into 30 pairs. Within each pair, one animal was randomly assigned to be in a “self-selection” condition in which the animals had some ability to choose what type and how much of a food or nutrient they ate per day. Each of the remaining 30 animals was randomly assigned to be paired (conceptually, not physically) with one of the animals in the self-selection (S) group. The second paired set of animals was called the randomization (R) group. Each animal in the R group was fed the diet its S partner ate the day before. The R animals had no choice but to adhere to those dietary amounts and proportions. The self-selection group received three different types of food *ad libitum* (F07171, 190 mg rodent purified diet with high fat, brown color, and chocolate flavor; F07172, 190 mg rodent purified diet with low fat, high carbohydrate, red color, and bacon flavor; and F07173, 190 mg grain-based diets with green color and banana flavor) and the forced choice pair got what the free choice one ate the day before. During the study all animals were fed once per day between 5 and 7 pm and were weighed biweekly. Sample size was calculated to achieve a power of 80% to detect a small to medium effect size with a 2-tailed test at an alpha level of 0.05.

Outcomes and covariates—At the end of the 2-week feeding regimen all CD1 mice were assessed in the open field and zero maze tests, after which they were assessed for body composition via quantitative magnetic resonance (QMR). Weight gain in grams, body composition (fat mass proportion and lean mass proportion), feed efficiency (calculated by dividing weight gain by total food consumption and then multiplying by 100), and activity measures (distances, speeds, and times spent in different areas in both the open field and zero maze tests) were considered the outcome variables. For each animal, the baseline body weight and the amount and proportion of diets they ate were recorded for the analysis.

Statistical analysis—The characteristics of mice in different groups were summarized as descriptive statistics (mean and standard deviation) and compared by t-test. The within-group comparisons were tested by paired t-test. In the weight gain study, general linear regression models were conducted to evaluate the association of food consumption with the weight gains in each group, and a linear mixed regression model was used to evaluate the group difference. In the R group, the food assignment was used for the food assumption, which was akin to an “intent-to-treat” analysis to keep the randomization intact.

Results

Study 1: Longevity and food consumption in mice

Descriptive statistics for the three groups (EO, OWL, and OWLM) are provided in Table 1. There were no statistical differences in baseline weight or energy intake before randomization between groups. There was a significant positive *association* between lifespan (birth to all-cause death) and self-selected daily energy intake for mice in each group

($P < 0.001$ (OWL), < 0.001 (OWLM), $= 0.009$ (EO) Figure 2), suggesting that 1 kcal more of daily energy intake was associated with an approximately 119.09 (OWL), 120.49 (OWLM) and 29.99 (EO) days longer lifespan, respectively. However, the result from the ANOVA for comparing the group means of lifespan revealed a significant difference between the groups ($P < 0.001$, Panel A in Figure 2), and the pairwise t-test revealed that the mean lifespan of the OWL group was the longest, followed by the OWLM and EO groups, indicating a negative association between calorie intake and lifespan.

To reduce the potential impact of reduced food intake in the moribund phase of the lifespan, daily energy intake was also calculated by the average over 120 days from the beginning of randomization and over the total number of days from the beginning of randomization to 1 week before death. We also conducted regressions excluding animals that were euthanized or that died accidentally because of technical reason (anesthesia) to account for censoring. The results from each of the sensitivity analyses showed similar results as above, that is, calorie intake had a negative effect on lifespan among groups and a positive association with lifespan within each of the three groups (data not shown).

Study 2: Weight gain and food consumption in young mice

The baseline body weight, food consumption, and weight gain of the mice in the 2-week feeding regimen are listed in Table 2. Mice in the S group had a larger mean body weight than mice in the R group at both baseline (38.2 vs 35.3, $P = 0.008$) and the study endpoint (42.9 vs 40.1, $P = 0.028$). However, there was no significant difference in weight gain (4.7 vs 4.8, $P = 0.961$) between the two groups. Although the diets consumed by each pair of mice were similar in type and amount, disparate associations were observed between the S and R groups. For animals in the S group, total food consumption was significantly associated with weight gain ($P = 0.001$) after control for baseline body weight (Table 3). Specifically, consumption of 1 more diet pellet (190 mg, or 0.64 kcal) by animals in the S group was associated with a 0.02 g greater weight gain on average. However, no significant causal effect of total food assignment on weight gain was detected in the R group. A linear mixed regression model showed a significant group-by-food consumption interaction for the association between weight gain and total food consumption ($P = 0.003$), indicating that the observational association estimates in the S group and the experimental effect estimates in the R group were significantly different (Table 4). The anxiety status of the mice was measured by their activities in both open field and zero maze tests; no significant differences were observed between the two groups (Table 5).

Discussion

In both studies we observed markedly disparate results of the observational association estimates and the experimental effect estimates regarding food consumption effects on lifespan and weight gain. With the randomized experimental design as the gold standard of causal inference, the observational association estimates in study 1 showed the opposite direction of the “true effect” of daily energy intake on lifespan in mice, whereas those estimates in study 2 did not represent the “true effect” of assigned food consumption on weight gain in young mice. Even we employed a study design using genetically identical

mice in the same environment; it is very possible that there are some unmeasured confounders not controlled in our observational designs, such as undiagnosed disease or individual characteristics, which can be strongly correlated to both exposure and outcomes and eventually bias the statistical inferences. For example, mice with any undiagnosed diseases or specific metabolic characteristics (e.g., basal metabolic rate, daily energy expenditure) may eat poorly and die earlier than the other healthy mice, which would produce a biased result. These two studies are typical examples of un-controlled confounding in studies with self-selection feature, which is frequently involved in food consumption research. The un-controlled confounders in self-selected food consumption (or other factors more generally) apparently can result in biased inference. Thus, relying on self-selection can lead to biased estimation of causal effects.

We acknowledge there are some limitations in this study. The study 1 was conceived shortly after the large project (NIA R01AG033682) had begun. We realized that the data to be generated could also be used to evaluate the associations of longevity and calories restriction under different designs and planned the analyses before the final data were collected. In Study 1, the mice are assigned (in the randomized controlled trial component) to conditions which differ in the upper limits of how much of a particular food they can eat. However, the mice in each group definitely can consume different amounts of food. Although we conclude in Study 1 that assignment to eat certain amounts of specific foods yields a different causal effect for the ingestion of those foods relative to the association with self-selection of ingestion of those foods, we acknowledge that the causal effect from the randomized groups is to the treatment assignment, which may not be the exactly same as to the amount of food eaten per se. However, when the intent-to-treat is applied, we may generally consider the group differences similar to the differences of the amount of food eaten. Additionally in Study 2, subjects in the randomized component voluntarily consumed approximately 9.1% less food than subjects in the nonrandomized component. It is possible that this total food intake interacted with and modified the effects of treatment assignment.

Some methods have been proposed to draw causal inferences from a statistical association in an observational study. One approach involves the use of a series of guidelines for conditions that must be met for an association to merit a conclusion of causation. The most well-known of these was offered by Sir Austin Bradford Hill [1]. However, the key to Hill's approach is that the association study alone never serves as sufficient justification for a conclusion of causation. Rather, a conclusion of causation is only tentatively drawn when multiple other conditions are met, several of which entail incorporating information from outside the association study.

The greatest challenge to drawing causal inferences in observational studies is the existence of potential confounding variables, not all of which can be specified, measured, or modeled. Randomization is the only method that can eliminate all potential confounders of the effect of treatment assignment per se, doing so by making the distribution of prerandomization factors identical for all treatment assignments at the population level [2]. In addition to the confounding issue, the assumptions of positivity and consistency for drawing causal inference are also hardly met in observation studies [16].

Multiple design and analysis procedures have been proposed for observational studies to address confounding. Madigan et al. evaluated the performances of many of these methods and concluded that “observational data can play an important role in the assessment of the effects of medical products, but no single analysis can provide definitive evidence” [17]. Can we remove the confounding bias in observational research by further improving the quality of study design, with more rigorous execution and appropriate statistical analysis? In our rodent feeding observational studies, all the mice shared a genetic background, were of the same age, were fed the same food, and lived in the same environment. Yet even in these meticulously designed and conducted studies, we still could not control all the confounders, and biased estimates of causal effects were observed. This applies *a fortiori* to studies in humans, where the inherent heterogeneity of human populations and the lack of knowledge of both biological and environmental confounders make it impossible in observational research to remove all potential confounders and, therefore, to draw unbiased causal inference.

Can the potential bias in causal estimates from observational research be rescued by statistical analysis strategies? Common statistical approaches to reducing bias in observational research are stratification, covariate adjustment, and propensity score methods [18]. However, all of these methods rely on the untestable assumptions that all confounders are included in the model [19], that the functional form of the relation between the measured confounders and the outcome or independent variable is modeled correctly [20], and that the confounders are measured without error [21]. Unfortunately, many important covariates (confounders) cannot be captured in observational studies because their identity is unknown or measuring them is infeasible.

Even though in theory discordance can occur between the association reported in an observational study and the causal effect estimated in a randomized design, it is an empirical question whether such discordance is large or common in practice. Some meta-analyses comparing RCTs and observational studies have found that randomized and observational studies yield similar findings in some situations [9, 10, 11, 12], but this is not always the case [13]. However, we must note that those studies were generally about medical treatments that physicians or other health care professionals apply to patients and not about nutrition and lifestyle factors that subjects self-select. The latter may be far more vulnerable to uncontrollable confounding. This is suggested by the results we have reported herein and also by meta-analyses in the nutrition field that showed discordance to the extent of opposite results between RCTs and observational studies [22, 23].

Is the discordance we observed likely the exception or the norm? That is difficult to say. Although we observed it in both of two cases, Shadish et al. [24] raised similar questions in the psychological domain, conducted a similar study in humans, and analyzed the data from the nonexperimental group using covariate adjustment and propensity score approaches. Shadish et al. found that concordant results could be obtained, but only with certain adjustment procedures. Whether the concordance Shadish et al. obtained versus the discordance we obtained represents idiosyncrasies of the situations studied or generalities of their educational setting versus our nutritional setting is unknown.

Observational designs remain useful in biomedical and behavioral research by allowing empirical investigations of exposures and their associations in populations experiencing all the vagaries of everyday life. Observational studies offer potential directions for randomized controlled experimental studies. However, as we have illustrated here, even an observational study that is meticulously controlled far beyond what could be achieved in a human study cannot be counted upon to reliably estimate causal effects owing to uncontrolled confounders, especially in nutrition research. Therefore, we believe that, despite public statements to the contrary [6], observational studies alone, no matter how well done, cannot support conclusions of causation.

Acknowledgments

This project was supported in part by NIH grants R01AG033682, P30DK056336 & P30AG050886. The funding bodies were not involved in the collection, analysis and interpretation of data, the writing of the manuscript, or the decision to submit for publication. DBA conceived the study idea and IK, TVG, TRN, DLS, and YY conducted the experiment. KE, PL, TRN, DLS, YY, AP, and DBA implemented statistical analysis. KE, PL, DLS, AP, and DBA jointly drafted the manuscript.

References

1. Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 1965; 58:295–300. [PubMed: 14283879]
2. Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics.* 1991; 47:1213–34. [PubMed: 1786315]
3. Cofield SS, Corona RV, Allison DB. Use of causal language in observational studies of obesity and nutrition. *Obes Facts.* 2010; 3:353–6. [PubMed: 21196788]
4. Bleske-Rechek A, Morrison KM, Heidtke LD. Causal inference from descriptions of experimental and non-experimental research: public understanding of correlation-versus-causation. *J Gen Psychol.* 2015; 142:48–70. [PubMed: 25539186]
5. Wang MT, Bolland MJ, Grey A. Reporting of limitations of observational research. *JAMA Intern Med.* 2015; 175:1571–2. [PubMed: 26053740]
6. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR good research practices task force report. *Value Health.* 2012; 15:217–30. [PubMed: 22433752]
7. Cole GD, Francis DP. Trials are best, ignore the rest: safety and efficacy of digoxin. *BMJ.* 2015; 351:h4662. [PubMed: 26320880]
8. Young SS, Karr A. Deming, data and observational studies: A process out of control and needing fixing. *Significance.* 2011; 8:116–20.
9. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000; 342:1878–86. [PubMed: 10861324]
10. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. *PLoS Med.* 2011; 8:e1001026. [PubMed: 21559325]
11. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000; 342:1887–92. [PubMed: 10861325]
12. Anglemeyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev.* 2014; 4:MR000034.
13. Ziff OJ, Lane DA, Samra M, Griffith M, Kirchhof P, Lip GY, et al. Safety and efficacy of digoxin: systematic review and meta-analysis of observational and controlled trial data. *BMJ.* 2015; 351:h4451. [PubMed: 26321114]

14. Vandenbroucke J. When are observational studies as credible as randomised trials? *Lancet*. 2004; 363:1728–31. [PubMed: 15158638]
15. Golub RM, Fontanarosa PB. Researchers, readers, and reporting guidelines: writing between the lines. *JAMA*. 2015; 313:1625–6. [PubMed: 25919525]
16. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes*. 2008; 32:S8–S14.
17. Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage JM, Suchard MA, et al. A systematic statistical approach to evaluating evidence from observational studies. *Annu Rev Stat Appl*. 2014; 1:11–39.
18. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*. 1997; 127:757–63. [PubMed: 9382394]
19. Spanos A. Revisiting the omitted variables argument: substantive vs. statistical adequacy. *Journal of Economic Methodology*. 2006; 13:179–218.
20. Becher H. The concept of residual confounding in regression models and some applications. *Stat Med*. 1992; 11:1747–58. [PubMed: 1485057]
21. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998; 55:651–6. [PubMed: 9930084]
22. Trikalinos TA, Moorthy D, Chung M, Yu WW, Lee J, Lichtenstein AH, et al. Concordance of randomized and nonrandomized studies was unrelated to translational patterns of two nutrient-disease associations. *J Clin Epidemiol*. 2012; 65:16–29. [PubMed: 22047889]
23. Miller PE, Perez V. Low-calorie sweeteners and body weight and composition: a meta-analysis of randomized controlled trials and prospective cohort studies. *Am J Clin Nutr*. 2014; 100:765–77. [PubMed: 24944060]
24. Shadish WR, Clark MH, Steiner PM. Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association*. 2008; 103:1334–43.

*based on body weight of animals alive at 10 months of age. Animals with overt health complications (e.g. active skin lesions, n=6 female, n=3 male) or within the lowest 1/3 body weight were culled prior to randomization.

**based on body weight of animals alive at 10 months of age. Animals with overt health complications (e.g. active skin lesions, n=5 males) within the heaviest 2/3 were also culled prior to randomization.

Figure 1.

Flow diagram of study 1. BW, body weight.

Figure 2.

Opposite association of food intake and lifespan in randomized controlled trial and observational study. (A): average lifespans in three diet groups. EO, ever obese group, in which mice were fed ad libitum; OWL, obese weight losers group, in which energy intake was restricted by nearly 30% of EO intake; OWLM, obese weight losers moderate group, in which energy intake was restricted by nearly 20% of EO intake. Error bar shows the standard error of mean intake. The P-values are from pairwise t-test with Bonferroni correction. (B, C, D): association of lifespan and daily energy intake for the mice in the (B) OWL, (C) OWLM, (D) EO group.

Table 1

Descriptive characteristics of mice included in the analysis.

Variables	OWL	OWLM	EO
N	80 *	170	78 **
Male	61	127	59 **
Wave 1	20	42	18 **
Wave 2	41	85	41
Female	19 *	43	19 **
Wave 1	19	43	19 **
Wave 2	0	0	0
Baseline weight (g)	44.23 (8.45)	44.10 (8.08)	44.48 (8.58)
Energy intake before randomization (kcal/day)	12.66 (1.66)	12.53 (1.72)	12.88 (1.55)
Mean energy intake (kcal/day)	10.80 (0.54)	11.79 (0.70)	14.34 (1.43)
Lifespan (days)	810 (156)	733 (162)	645 (145)

Note: Baseline weight, intake before randomization, mean energy intake, and lifespan are shown as mean (standard error). EO, ever obese group, in which mice were fed ad libitum; OWL, obese weight losers group, in which energy intake was restricted by nearly 30% of EO intake; OWLM, obese weight losers moderate group, in which energy intake was restricted by nearly 20% of EO intake.

* One female mouse is excluded from the analysis because it died immediately after the randomization preventing food intake determination.

** One male and two female mice are excluded from the analysis as they were euthanized for tissue collection rather than observed death or moribund status.

Table 2

Food consumption and weight gain of the mice.

Variables	Self-Selection Group	Randomization Group	P values [*]
	Mean (SD)	Mean (SD)	
N	30	30	-
Pre-randomization pellets eaten	158.8 (16.7)	162.4 (16.6)	0.401
Baseline BW (g)	38.2 (3.7)	35.3 (4.1)	0.008
Endpoint BW (g)	42.9 (4.9)	40.1 (4.7)	0.028
Weight gain (g)	4.7 (1.6)	4.8 (1.7)	0.961
Banana flavor ^{\$}	39.3 (23.2) [#]	29.8 (12.2) [#]	0.052
Chocolate flavor ^{\$}	185.2 (35.8)	164.9 (34.7)	0.030
Bacon flavor ^{\$}	189.5 (28.7)	181.4 (28.2)	0.273
Total food ^{\$}	414.0 (48.6)	376.0 (39.5)	0.002
Banana-flavor proportion ^{\$}	0.09 (0.05)	0.08 (0.03)	0.222
Chocolate-flavor proportion ^{\$}	0.45 (0.07)	0.44 (0.08)	0.312
Bacon-flavor proportion ^{\$}	0.46 (0.07)	0.48 (0.07)	0.203

Note: BW, body weight; SD, standard deviation.

^{\$}The diets consumed in the self-selection group were also the diets assigned to the randomization group.^{*} Student's t-test.[#] P<0.0001 compared to chocolate consumed or bacon consumed, by paired t-test.

Table 3

Disparate associations of weight gain and food consumption in the self-selection and randomization groups by general linear regression models.

Parameter	Self-Selection Group		Randomization Group	
	Estimate (SE)	P values *	Estimate (SE)	P values *
Intercept	−8.20 (2.12)	0.001	3.29 (3.26)	0.322
Baseline weight	0.13 (0.07)	0.057	0.07 (0.09)	0.449
Total food	0.02 (0.01)	0.001	−0.003 (0.009)	0.746

Note: SE, standard error.

* Likelihood ratio F test.

Table 4

Group-specific association of weight gain and food consumption by a general linear mixed regression model.

Parameter	Estimate	SE	P values
Intercept	-7.60	2.58	0.006
Group (R=1; S=0)	10.41	3.08	0.002
Baseline weight	0.10	0.06	0.099
Total food	0.02	0.01	0.002
Group*Total food	-0.02	0.01	0.003

Note: R, randomization group; S, self-selection group; SE, standard error.

Table 5

Activities of mice in the open field and zero maze tests.

Variables	Self-Selection Group	Randomization Group	P values *
	Mean (SD)	Mean (SD)	
N	30	30	-
Open Field Test			
Distance (cm)	2221.8 (491.8)	2274.9 (582.4)	0.704
Speed (cm/s)	9.3 (2.0)	9.5 (2.4)	0.704
Time at center (s)	15.9 (11.9)	13.9 (8.0)	0.430
Time at side (s)	224.1 (11.9)	226.2 (8.0)	0.429
Zero Maze Test			
Distance (cm)	964.7 (185.7)	959.0 (875.0)	0.914
Speed (cm/s)	4.0 (0.8)	4.0 (0.9)	0.917
Time at open area (s)	64.2 (23.1)	63.8 (54.4)	0.946
Time at closed area (s)	176.0 (23.1)	176.5 (25.0)	0.946

Note: SD, standard deviation.

* Student's t-test.