Toward more rigorous and informative nutritional epidemiology: The rational space between dismissal and defense of the status quo

Andrew W. Brown, Stella Aslibekyan, Dennis Bier, Rafael Ferreira da Silva, Adam Hoover, David M. Klurfeld, Eric Loken, Evan Mayo-Wilson, Nir Menachemi, Greg Pavela, Patrick D., Quinn, Dale Schoeller, Carmen, Tekwe, Danny, Valdez, Colby J. Vorland, Leah D. Whigham & David B. Allison

Published online: 22 Oct 2021.

Submit your article to this journal ⬀

Article views: 1290

View related articles ⬀

View Crossmark data ⬀

Taylor & Francis
Taylor & Francis Group

REVIEW

🔓 OPEN ACCESS | Check for updates

Toward more rigorous and informative nutritional epidemiology: The rational space between dismissal and defense of the status quo

Andrew W. Brown[a] 🔟, Stella Aslibekyan[b,c], Dennis Bier[d], Rafael Ferreira da Silva[e], Adam Hoover[f], David M. Klurfeld[g] 🔟, Eric Loken[h], Evan Mayo-Wilson[i] 🔟, Nir Menachemi[j], Greg Pavela[k], Patrick D. Quinn[a] 🔟, Dale Schoeller[l], Carmen, Tekwe[i] 🔟, Danny, Valdez[a] 🔟, Colby J. Vorland[a] 🔟, Leah D. Whigham[m], and David B. Allison[i] 🔟

[a]Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, Indiana, USA; [b]Department of Epidemiology, University of Alabama at Birmingham, Brimingham, Alabama, USA; [c]Department of Epidemiology, College of Public Health, University of Kentucky, Lexington, Kentucky, USA; [d]Children's Nutrition Research Center, Baylor College of Medicine, Houston, Texas, USA; [e]Department of Computer Science, Viterbi School of Engineering, University of Southern California, Los Angeles, California, USA; [f]Holcombe Department of Electrical and Computer Engineering, The College of Engineering, Computing and Applied Sciences, Clemson University, Clemson, South Carolina, USA; [g]Agricultural Research Service, United States Department of Agriculture, Beltsville, Maryland, USA; [h]Department of Educational Psychology, Neag School of Education, University of Connecticut, Storrs, Connecticut, USA; [i]Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, Indiana, USA; [j]Department of Health Policy and Management, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, Indiana, USA; [k]Department of Health Behavior, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, USA; [l]Department of Nutritional Sciences, University of Wisconsin-Madison Biotechnology Center, Madison, Wisconsin, USA; [m]Center for Community Health Impact and Department of Health Promotion and Behavioral Sciences, University of Texas Health Science Center School of Public Health, El Paso, Texas, USA

ABSTRACT

To date, nutritional epidemiology has relied heavily on relatively weak methods including simple observational designs and substandard measurements. Despite low internal validity and other sources of bias, claims of causality are made commonly in this literature. Nutritional epidemiology investigations can be improved through greater scientific rigor and adherence to scientific reporting commensurate with research methods used. Some commentators advocate jettisoning nutritional epidemiology entirely, perhaps believing improvements are impossible. Still others support only normative refinements. But neither abolition nor minor tweaks are appropriate. Nutritional epidemiology, in its present state, offers utility, yet also needs marked, reformational renovation. Changing the status quo will require ongoing, unflinching scrutiny of research questions, practices, and reporting—and a willingness to admit that "good enough" is no longer good enough. As such, a workshop entitled "Toward more rigorous and informative nutritional epidemiology: the rational space between dismissal and defense of the status quo" was held from July 15 to August 14, 2020. This virtual symposium focused on: (1) Stronger Designs, (2) Stronger Measurement, (3) Stronger Analyses, and (4) Stronger Execution and Reporting. Participants from several leading academic institutions explored existing, evolving, and new better practices, tools, and techniques to collaboratively advance specific recommendations for strengthening nutritional epidemiology.

Main text introduction

Introduction

Nutritional epidemiology has been characterized by some critics in extreme terms as *absurd*, *corrupt*, and even *a dead science*. In *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth*, Stuart Ritchie (2020) noted:

> Fads like microbiome mania wax and wane, but there's one field of research that consistently generates more hype, inspires more media interest and suffers more from the deficiencies outlined in this book than any other. It is, of course, nutrition. The media has a ravenous appetite for its supposed findings:

> 'The Scary New Science That Shows Milk is Bad for You'; 'Killer Full English: Bacon Ups Cancer Risk'; 'New Study Finds Eggs Will Break Your Heart'. Given the sheer volume of coverage, and the number of conflicting assertions about how we should change our diets, little wonder the public are confused about what they should be eating. After years of exaggerated findings, the public now lacks confidence and is sceptical of the field's research. (Ritchie 2020)

Such skepticism is not relegated to the public. Via a commentary in *The BMJ*, Nina Teicholz and Gary Taubes (2018) maintained:

> Despite methodological advances, nutritional epidemiology remains fundamentally limited by its observational nature.

Guidelines relying on the circumstantial evidence can be a little more than educated guesses. (Teicholz and Taubes 2018)

Some academicians even suggest abolishing nutritional epidemiology. As Scott Lear, a professor of Health Sciences at Simon Fraser University, posited, "One may wonder if we should stop nutritional research altogether until we can get it right" (Lear 2019). John P.A. Ioannidis stated, "Nutrition epidemiology is a field that's grown old and died. At some point, we need to bury the corpse and move on to a more open, transparent sharing and controlled experimental way" (Belluz 2018).

But the field—and the status quo—also has its defenders. Ambika Satija and colleagues asserted in *Advances in Nutrition* (Satija et al. 2015):

Nutritional epidemiology has recently been criticized on several fronts, including the inability to measure diet accurately, and for its reliance on observational studies to address etiologic questions.… These criticisms, to a large degree, stem from a misunderstanding of the methodologic issues.… Misunderstanding these issues can lead to the non-constructive and sometimes naïve criticisms we see today. (Satija et al. 2015)

Rightly or wrongly, nutritional epidemiology's research findings have played a large role in shaping how we perceive relationships between food or nutrients and disease and how national policy guidelines are determined. These research findings also shape public opinion and affect public health. But nutritional epidemiology's research outcomes are all too often derived by study modalities that yield only low-level evidence.

Despite limitations in the study designs commonly used—not to mention poor executions of analyses and misreporting of subsequent results—researchers often make claims of causality when reporting nutritional associations (Cofield, Corona, and Allison 2010). In particular, the discipline has relied heavily on simple observational studies and meta-analyses of these simple observational studies. But these ordinary association tests alone cannot determine causality. At best, simple observational studies may, as part of a larger body of evidence, result in collective evidence of causation sufficient for some standards (Hill 1965).

Moreover, current methods of measuring dietary intake, food composition, and environmental "exposome" covariates arguably fall short of both the accuracy and the precision necessary to confidently detect causal risk relationships or their magnitude, and do not meet the standards of quality often held in other research domains (Schoeller and Westerterp 2017; Patel and Ioannidis 2014).

Nutritional epidemiology can—and must—do better by pursuing greater scientific rigor, academic honesty, and intellectual integrity. And the time is right to do so. Some academicians, believing such change is impossible, wish to jettison the tools upon which nutritional epidemiologists historically have relied. Still others advocate for only normative refinements. But neither abolition nor minor tweaks are appropriate. Nutritional epidemiology, in its present state, offers utility and substantial room for improvement. To change the status quo will require an ongoing, collective examination of nutritional epidemiology's research questions, practices, and reporting—and a willingness to admit that "good enough" is no longer good enough.

In the spirit of strengthening the field of nutritional epidemiology, an online event, "Toward more rigorous and informative nutritional epidemiology: the rational space between dismissal and defense of the status quo," was held from July 15 to August 14, 2020. The symposium comprised 15 prepared research talks, several moderated panel discussions, and small-group, open-forum sessions related to the need for reforms in four areas of focus: (1) Stronger Designs, (2) Stronger Measurement, (3) Stronger Analyses, and (4) Stronger Execution and Reporting.

Invited participants from several leading academic institutions explored new best practices, tools, and techniques to strengthen the nutritional epidemiology field. Following small-group discussions, the working groups presented their findings, considered the various perspectives offered, and then collaboratively worked through specific recommendations. For each of the four focus areas, we first summarize the recommendations that resulted from the prepared talks and discussions. We then summarize and expand on the content of the prepared talks and discussions. Finally, we provide some concluding comments.

Points considered

Stronger designs

Recommendations

- Begin with the research question to be answered, consider which measurements could most effectively answer this question, and develop a study design best-suited to delivering these data. Researchers can strengthen the quality of observational research by expanding study design options beyond traditional observational methods or by combining traditional observational methods with more objective measurements.
- First consider what a design can and cannot accomplish (to the extent this is known). Such consideration is essential for understanding a study design's strengths and limitations and systematically addressing any assumptions and limitations, such as through sensitivity analyses, falsification tests, or constraints on conclusions.
- Designs exist both for generalizability of results and for making inferences to a specific individual (e.g., pragmatic trials versus repeated N-of-1 trials, respectively), and investigators should be precise about the population or individual to which inferences are appropriately made.
- Before declaring a randomized trial to be impossible, impractical, or unethical, researchers should thoroughly review all available options. Myriad design options are employed across academic disciplines, including conventional trials, unconventional trials, and emerging approaches. Creative solutions may

make a randomized design possible, practical, and ethical, thereby adding stronger causal inference to some nutritional epidemiological questions.

Discussion summary

Beyond the dichotomy of ordinary association tests versus randomization. Much of the nutritional epidemiological literature has relied on simple observational studies – specifically ordinary association tests (OATs) These have been defined as:

> Observational studies on samples of individuals in which the sole or primary means of controlling for potential confounding factors is inclusion of measures of some potential confounding factors as covariates in statistical models (or stratifying by measures of such factors). OATs are heavily relied upon in thinking about plausible effects of policies, but have also been heavily criticized in general and in the obesity and nutrition domains in particular for multiple reasons. (Richardson et al. 2017)

The limitations of OATs as a means of reliably determining causation are well established (Jepsen et al. 2004), and these limitations are not necessarily specific to nutritional epidemiology. Indeed, Hernán, in discussing the use of causal language for epidemiological research, succinctly titled a section, "Of course 'association is not causation,'" and outlines the importance of articulating better causal questions than those calculated by simple associations (Hernán 2018). Such tests *do*, however, have some merit. At a minimum, they assist in the generation of hypotheses—even if OATs, themselves, are poor tests of the hypotheses they help to generate. But unless a study is designed to *strengthen* an inference—and not simply to rehash others' findings—it will do little to advance nutritional epidemiology's body of knowledge and may merely amplify bias. The literature remains replete with OATs, sometimes repeating different iterations of the same association dozens upon dozens of times (e.g., Brown, Bohan Brown, and Allison 2013). The field needs to use more robust modes of inquiry, some of which we discuss below, especially when it becomes clear that yet another OAT will not advance the causal—or even associational—understanding of a nutrition-health relation (Brown et al. 2014).

While randomization may be the gold standard, in some cases it is not feasible or ethical. As *Randomistas* author Andrew Leigh (2018) explains:

> Not every intervention can—or should—be randomized. A famous article in *The British Medical Journal* searched the literature for randomized trials of parachute effectiveness. Finding none, the researchers concluded (tongue-in-cheek): 'the apparent protective effect of parachutes may be merely an example of the "healthy cohort" effect… The widespread use of parachutes may just be another example of doctors' obsession with disease prevention.' Using similar critiques to those leveled at non-randomized studies in other fields, the article pointed out the absurdity of expecting everything to be randomly trialed. (Leigh 2018)

Others have similarly outlined the challenges to implementing conventional randomized designs for questions related to diet (Hébert et al. 2016). The parachute analogy, or those like it at the logical extreme (Katz 2019), is not directly comparable to questions of how nutrition relates to chronic disease. Nutrition can impact chronic disease outcomes in multiple ways and with generally small effect sizes. A lack of a parachute, by contrast, has one causal pathway and large, clearly observed acute effects (Hayes et al. 2018). Nonetheless, other biomedical disciplines have falsely claimed that their interventions were akin to a parachute, even when randomized trials were conducted to test the intervention (Hayes et al. 2018).

Still, challenges with applying randomization need not justify acceptance of the status quo. Modifications can be made. As Leigh continues, "The parachute study has been widely quoted by critics of randomized evaluations. But it turns out that experiments on parachute efficacy and safety are widespread" (Leigh 2018). Crash test dummies have been used in impact testing of jumps from various altitudes, and paratroopers were randomized to protective ankle braces which were found to reduce parachuting-related ankle sprains by a factor of six (Leigh 2018).

There is important and worthwhile middle ground between the position that conventional randomization is the *only* valid avenue and the notion that, for cases in which conventional randomization is not possible, *any* nonrandomized study is equally acceptable and valid. In actuality, a mix of conventional and unconventional interventional approaches and quasi-experimental designs can be leveraged to great effect. As such, it is time for scholars to broaden their research modalities and to employ creativity during a study's initial design.

Researchers also need to stay current on the growing number of available novel designs when simpler designs cannot answer the research question, and understand how to appropriately select, execute, and analyze results from them. Many textbooks on clinical and experimental research design in the behavioral health sciences emphasize "conventional" experimental or quasi-experimental designs, including (but not limited to) two-group parallel arm, factorial, withdrawal, crossover, and pragmatic trial designs (Friedman et al. 2010; Shadish, Campbell, and Cook 2001; Windsor et al. 2001; Meinert 1986). While such designs are essential to the experimentalist's toolkit, novel or less well-known variations on these designs may be useful. Stepped-wedge cluster randomized designs (Hemming et al. 2015); within-cohort randomized trials (also called Trials within Cohorts) (Kim, Flory, and Relton 2018); Randomization to Randomization (R2R) (George et al. 2018); packet-randomized experiments (Pavela et al. 2015); multiphase optimization strategy trials (MOST) (Collins 2018); sequential, multiple assignment, randomized trials (SMART) (Almirall et al. 2014; Lei et al. 2012); and repeated N-of-1 trials (Duan, Kravitz, and Schmid 2013) each extend or provide alternatives to the standard randomized design by varying design features including the planned timing of treatment assignment, treatment optimization criteria, participant expectancies, the consent process, and nature of the treatment. Increased familiarity with these designs and other novel designs, as well as the reasons for using them, will make it more likely that researchers will select an appropriate randomized design.

The question remains as to which randomized designs provide the most relevant evidence for health care providers versus policymakers. A particularly salient example may be what Sacristán and Dilla (2018) call the "contradiction" of the pragmatic design:

> The root of the contradiction is that the same model that considers that a pragmatic attitude aims to inform clinical decision-making assumes that health care decision-makers speak the language of populations. In reality, while historically decisions made by policy-makers have been population based, clinical decisions are always individual based. (Sacristán and Dilla 2018)

This "contradiction" has become more evident with growing interest in precision-medicine and the possibility of making inferences to individuals rather than to populations by using N-of-1 trials, which resemble the conventional crossover design in that they are multiple-period crossover experiments comparing two or more treatments, but within individual patients. It is encouraging to see debates about the relevance of different randomized designs to decision-making (Pavela et al. 2015).

Nonrandomized studies, too, have their place—provided that their designs' potential limitations and assumptions are systematically evaluated and addressed accordingly. Furthermore, not all nonrandomized designs are the same, ranging from controlled (but nonrandomized) interventions to OATs. At a minimum, a thoughtful consideration regarding the goal of nonrandomized research, such as causal inference, should be articulated (c.f., Chiu et al. 2021; Tobias and Lajous 2021). Error can influence study results in any direction, and even some stronger nonrandomized designs may, in practice, inadvertently exacerbate some bias.

Example: Family-based designs. Family-based designs, like all designs, are susceptible to threats to internal validity and have other assumptions that may or may not be met in a given application. To compare outcomes among siblings who do and do not experience an exposure or intervention, for example, a family-based design exploits familial relatedness to enhance confounder control. Yet, the design does not obviate the need for longitudinal data, measured at a suitable timescale, to rule out reverse causation and provide an appropriate test of the hypothesized effect (McGue, Osler, and Christensen 2010). Neither will the design, in and of itself, resolve bias from nonrandom measurement error (Trepanowski and Ioannidis 2018). Moreover, the design can be especially vulnerable to bias from random measurement error and unmeasured confounders that are not shared by family members (Frisell et al. 2012). Further limitations stem mostly from the requirement of discordance—the use of within-family variation in exposures and outcomes to estimate associations (D'Onofrio et al. 2016).

Even so, sibling comparisons—especially comparisons of discordant twins—were applied successfully as early as the late 18th century. Considered a health hazard at the time, coffee consumption was banned in Sweden. King Gustav III ordered a study on a pair of identical twins:

> One twin agreed to drink three pots of coffee for the rest of his life, and the other one a similar amount of tea. Two prominent physicians were monitoring their health. Both physicians died before the experiment completed, one dying before the other. Gustav III himself was assassinated in 1792, while both twins lived healthily for a long time. Eventually, the tea consumer twin died at the age 83 years, and coffee won! (Afshari 2017)

Sweden's coffee ban would be reversed in the 1820s, but today science demands stronger evidence than comparisons between only two twins. By the mid-20th century, some epidemiological uses of twin studies would include ruling out genetic confounding in associations of tobacco smoking with mortality (Lichtenstein et al. 2002; Kaprio and Koskenvuo 1989), studying correlates of obesity in small samples of well-characterized discordant twins and diet and mortality in larger samples linked to national health registers (Naukkarinen et al. 2012; Granic et al. 2013), and exploring the association of exposure to breastfeeding with obesity in childhood and adolescence using multiple sibling comparisons (Metzger and McDade 2010; Colen and Ramey 2014). There is potential for extending family-based designs to omics research—for instance, examining identical twins discordant for dietary factors (Pallister, Spector, and Menni 2014; Barron et al. 2016).

While family-based approaches can enable researchers to control for sources of familial similarity (e.g., genetic factors, educational background, home environment, parenting practices) *by design* (i.e., without the need for measured covariates), the benefits of such study designs to nutritional epidemiology are bound by some conditions. Family-based designs are likely to be used to greatest effect in nutritional epidemiological study when (1) constructs, such as nutritional exposures, can be measured well, (2) unmeasured confounders are likely shared among siblings or other family members, (3) exposures vary within the family members studied (twins, siblings, etc.), and (4) limitations and assumption violations can be identified and assessed, including through the use of multiple family-based designs and systematic sensitivity analyses (D'Onofrio et al. 2016).

Challenges moving forward. No matter which study designs they use, researchers should be willing to tolerate and openly share statements of uncertainty and to publish their results, regardless of characteristics like statistical significance or consistency with the present zeitgeist. Unfortunately, doing so can present challenges for publishing results.

After conducting a systematic review of the literature to examine the relationship between built environments and

physical activity or obesity rates, a group of researchers (including authors on this paper) identified the need for higher-quality evidence, noting:

> Recognizing that experimental studies are potentially not feasible in many situations, researchers should look for opportunities to employ quasi-experimental designs. One example of such designs is the difference-in-difference approach that seems particularly applicable for the study of environmental changes such as the addition of a greenway to a neighborhood. (Ferdinand et al. 2012)

Upon learning that a new park would be erected in downtown Birmingham, AL, USA, the same group of researchers decided to study its impact on the body mass index (BMI) of children living nearby (Goldsby et al. 2016). They extracted changes in BMI from electronic health records collected by downtown Birmingham clinics and tested whether children living closer to the new park exhibited changes in BMI, pre- and post-park, relative to those of children who lived farther away from the park.

Using difference-in-difference statistical modeling, they investigated park proximity and its associations on children's BMI. The main takeaway? "Proximity to a park was not associated with reductions in BMI z-score" (Goldsby et al. 2016).

The researchers were forthcoming about their study's limitations:

> The sample sizes of the near groups were relatively small, potentially limiting the power to identify significant differences between groups. Having more children in the near groups would have been ideal, but being able to examine BMI longitudinally, even in a small group of children, provides valuable information for other obesity researchers and policymakers working to address the U.S. obesity epidemic. (Goldsby et al. 2016)

However, it took myriad attempts for the researchers to find a journal willing to publish their findings. This difficulty may have been in part because the results ran counter to conventional thinking about the potential health benefits of proximity to green spaces. It may also have been in part that, although common to other fields of study, the researchers' more rigorous, quasi-experimental approach was generally unfamiliar to reviewers.

Nonetheless, null results still offer value. According to *Reproducibility and Replicability in Science*, produced by the National Academies of Sciences, Engineering, and Medicine (NASEM), "The advent of new scientific knowledge that displaces or reframes previous knowledge should not be interpreted as a weakness in science" (2019). On the contrary, such occurrences are a function of science's "continuous process of refinement to uncover ever-closer approximations to the truth."

Rather than put forward yet another simple observational study, the group of researchers highlighted a different method, raised new questions, and created a roadmap for continued exploration. It is in these ways that the body of nutritional epidemiological knowledge can be made to move forward.

Stronger measurement

Recommendations

- Self-reporting tools have utility for some uses; however, when possible, self-report should be used in conjunction with additional, objective means of evidence validation, and should be avoided when invalid or unfit for a particular use. Increasing the accessibility of information about available objective measurements, their appropriate uses, and their relative costs could facilitate their increased use.
- Blending varying degrees of automation with traditional, observational studies can improve the quality of self-reported data. Ideally, investigators should have access to completely independent methods of determining food intakes, comprehensive analyses of the chemical compositions of foods that include ranges of nutrient variability, and fully independent methods of assessing specific nutrient intakes.
- Researchers should continually seek additional biomarkers and other new technologies and methods for collecting objective data. Multi-disciplinary partnerships and interactions may be able to hasten improvements in available objective measurements. Institutions and funders should prioritize the development, training, and use of such improvements.

Discussion summary

Status quo: Self-report. The field's continued reliance on substandard measurements has hindered progress in nutritional epidemiology. In particular, traditional observational methods such as self-reporting remain mainstays, despite their potential for inaccuracy.

Consider, for instance, the assessment of energy intake using self-reporting. Self-reported energy intake was first compared with doubly-labeled water—a biomarker of habitual energy intake—in 1986. The self-reported measure *underestimated* energy intake by 34% in women with obesity (Prentice et al. 1986).

Moreover, systematic reviews have since identified 59 studies, including 6,298 adults, and 15 studies in 664 children that compared energy intake from food diaries, 24-hour recalls, or food frequency questionnaires (FFQs) with doubly-labeled water energy expenditure (Burrows et al. 2019; Walker, Ardouin, and Burrows 2018). Underreporting of energy intake averaged about 20%, but varied from 1% to 67% across these studies. Underreporting was common in participants older than eight years, increased with body mass index, and was found in countries at all stages of economic development. Attempts to reduce bias in energy intake estimates by excluding extreme values, for example, the use of "Goldberg cutoffs," have been shown to be unreliable (Ejima et al. 2019). These problems with self-report-based estimates of energy

balance have resulted in calls to discontinue their use in the calculation of actual energy balance (Dhurandhar et al. 2015). Yet, their use continues.

Studies using biomarkers of protein, potassium, and sodium intake have found that underreporting occurs most for foods characterized by lower protein and sodium content, which is consistent with selective underreporting of high-fat, high-sodium snack and savory food items. These deficiencies were first reported over 35 years ago and have been confirmed in multiple studies. Yet, their use continues.

The FFQ, which is most commonly applied in large cohort studies, does not accurately estimate frequency of intake or gauge serving size (Willett et al. 1987). Correlations of intake data with a limited number of plasma biomarkers suggest mostly weak to moderate associations among large groups (Cade et al. 2004). Correction for energy intake might account for different energy needs. However, energy intake from FFQs is invalid (Dhurandhar et al. 2015), and thus correcting for it calls into question all extrapolated nutrient data, especially for extrapolation of risk for individuals (Krall and Dwyer 1987). There do not appear to be any modifications of the FFQ that would produce accurate and precise data (Kipnis et al. 2002). The clearest benefit of the use of FFQs is the collection of some degree of dietary information before disease occurs, precluding reverse causality by accounting for a temporal relationship.

Dietary intake assessments frequently depend on conscious recall of the foods being studied. The data on the *contents* of all of the nutrients and bioactive substances present in the reported foods being consumed are also a source of error. And so are the data on the *variability* of contents—due to individual cultivars, harvesting and storage conditions, food preparation and cooking methods, and the relationships of these to the lifestyle, behavioral, and environmental variables (sometimes referred to as an "exposome") that co-vary with dietary intakes.

Despite such well-documented but frequently unaddressed short comings, the use of self-reported dietary assessment instruments remains commonplace. Proponents of the continued use of self-reported nutritional data often surmise that using such methods is better than nothing, given the importance of diet in the maintenance of health and development of disease (Satija et al. 2015). Unfortunately, when used in isolation, self-reporting is a rather blunt instrument—one that can limit the scope of a study's design and negatively impact the nature and quality of the research questions pursued. Indeed, in the example of self-reported energy balance, the direct use of self-report has been worse than nothing (Dhurandhar et al. 2015).

Utility of self-report. That is not to say there is no longer any room for self-reporting. Two of the most common methods for assessing dietary intake, 24-hour recalls and FFQs, can be implemented in many ways, and some of these implementations—like the Automated Multiple Pass Method (AMPM) and the ASA24 (adapted version of AMPM for self-administration)—perform better than others (Moshfegh et al. 2008).

Originally developed by the U.S. Department of Agriculture (USDA)/Agricultural Research Service (ARS) for the National Health and Nutrition Examination Survey (NHANES), the AMPM puts energy intake within 3% of estimates from doubly-labeled water in people with BMI < 25 (Moshfegh et al. 2008). However, with increasing degree of overweight, this decreases to 80% of true intake, so overall population estimate is 89% of actual energy requirement. Additional validation of sodium intake, which correlates strongly with energy in the diet, showed 90% recovery in urine samples (Rhodes et al. 2013). Even these levels may not be adequate for usual weight changes but may be enough for macronutrients.

NHANES uses two nonconsecutive recalls to estimate intake. The first of these—a face-to-face meeting with trained personnel—uses three-dimensional models to estimate food serving sizes. The second interaction occurs via telephone and uses actual-size, two-dimensional food images in booklets provided to participants. It is well understood that older-style, 24-hour recalls must be administered multiple times to estimate nutrient intake in individuals, whereas a much smaller number of recalls is needed to approximate population averages, which is the proper use of NHANES data (Basiotis et al. 1987). NHANES data are cross-sectional, and thus provide weak causal evidence, even when repeated over time.

Because it takes 30 minutes to complete and is administered in the NHANES setting, the 24-hour AMPM may be ill-suited to large cohort studies due to time and cost. But the National Cancer Institute (NCI) has developed an internet-based, self-administered version of the AMPM called the Automated Self-Administered 24-hour (ASA24) Dietary Assessment Tool. ASA24 similarly may afford more standardization than other, traditional observational methods (Subar et al. 2012). Investigators can also rely on limited biomarkers for other nutrients to assess the performance of these diet assessment tools.

Moving beyond self-report in isolation. Combining traditional observational methods with more objective measurements can greatly boost a study's utility, but in deciding which objective measurements to include in a design, researchers across scientific disciplines face the same trilemma. The ideal tool would provide measurements that are 1) accurate and precise, 2) detailed, and 3) frequent. However, the fundamental nature of a trilemma is that it is impossible to secure all three equally and simultaneously.

This notion certainly applies to dietary measurements. Case in point: while food diaries, 24-hour recalls, and FFQs all focus on dietary detail, that emphasis on detail inherently reduces both the accuracy and frequency of measurement. On the other hand, a tool such as doubly-labeled water provides accurate measures of metabolizable energy, but little in the way of detail (Speakman et al. 2021). For some diet-disease relationships, knowledge of specific nutrient intake is important. However, for obesity, the most prevalent

diet-related condition in the United States, no clear dietary intervention prevails for long-term efficacy. For obesity, measurement tools need to focus on accuracy and frequency of measurement of energy intake.

Some of the tools nutritional epidemiologists can leverage to collect data more accurately and with more frequency include wearable devices developed to detect and measure consumption (Hoover and Sazonov 2016). These devices can reduce underestimation of energy intake by providing an objective measure that does not rely on self-reporting (Salley et al. 2016).

For example, in one study of automated bite counting that used wrist-motion tracking, estimates of energy intake were significantly more accurate than a guess, and automated bite counting was comparable to human estimates using a detailed menu. By automating the measurement process, these devices also reduce cognitive load and, hence, user burden (Weathers, Siemens, and Kopp 2017). This, in turn, helps to increase frequency of measurement.

Although these wearables may be designed to accurately estimate energy intake, they offer poor precision. Some wearables have poor precision because they treat all foods equally. One study of 30 subjects that used a sensor to measure chews and swallows, found an average error of approximately 30% of energy intake per validation meal and 16% for training meals, compared to investigator measured intake (Fontana et al. 2015). The study also used estimates from photographic food records, which showed approximately 20% error in both cases. Another study of 77 people compared automatic bite count with kilocalories measured by use of 24-hour recall over a two-week period (2,975 meals/snacks) and found a per-meal correlation of 0.53 (Scisco, Muth, and Hoover 2014). While the self-reported intake in the latter study was subject to the issues already discussed with self-report, in both of these studies, the average per-meal accuracy was high, at the expense of lower per-meal precision.

Although these examples focused on energy intake, other technologies exist (e.g., photogrammetric approach, continuous blood glucose monitoring, metabolomic profiles). Each case sacrifices parts of the trilemma. A bite counter gives (on average) accurate energy intake, details on consumption and chewing patterns, and frequent measurement, but lacks information on other characteristics of the food or meal. Photogrammetric approaches provide a richer detail of meal content and context, but do not provide the same information on eating rate and may require user input to calibrate information for specific meals. Despite their limitations, however, the newer technologic and biomarker approaches provide objective measurements that are not dependent on self-report.

Addressing nutritional epidemiology's data-related challenges may entail developing entirely new ways of measuring food consumption, assessing specific nutrient intakes, and of more accurately analyzing the chemical compositions of foods. Collecting data via a mix of more objective measurements and using increasingly robust study designs will help to advance nutritional epidemiology. But for these changes to be truly effective, additional reforms related to the analysis of these data are also needed.

Stronger analyses

Recommendations

- The relationship of dietary factors to numerous potential confounders, such as age, sex, education, and income, should be determined, and uniform standards developed to include and address these.
- Investigators should use multiple analytical methods, including appropriately robust and sometimes novel statistical tools, to mitigate biases common to simple observational studies.
- To resolve the complex problem of innumerable interacting variables in the exposome, investigators should seek information technology approaches to the investigation, reduction, and interpretation of data.

Discussion summary

The factors discussed in "Stronger Designs" and "Stronger Measurements" allow the field of nutritional epidemiology to employ strategies that result in stronger prediction and causation (Imbens and Rubin 2015; Rosenbaum and Rubin 1983; Morgan and Winship 2015). Yet many investigations still use OATs to explore the relationship between X and Y, in which an investigator might speculate that X *causes* Y. After observing the association of X and Y, a potential confounder, Z, arises. So the investigator measures and controls for the presumed confounder, Z. What is the problem with this method? In addition to the quality of inference depending critically on the quality of the measurement (discussed above), failure to model the functional form correctly can at best reduce and at worst contribute to bias (Westfall and Yarkoni 2016).

Reliance on just one approach for ruling out alternative explanations can leave invalidities or biases behind. When *multiple* methods are incorporated, they can greatly reduce the number of alternative explanations. In this section, we touch on three potential concerns for inference: within- versus between-subject effects, incorporating complexities of new measurement methods, and stronger inferential analyses.

Within- versus between-subject effects. Analysis of "between-person" and "within-person" data can provide clarity to research questions. To the extent observational data are used to support causal inference, assumptions must be considered for how X (the independent variable) was "assigned" in the population, such as the *level of analysis* at which the variables are covarying.

Often, X, Y, and Z data are collected on a sample of people and, so, the covariation matrix represents the *between-person* covariation. However, if the variables were

collected as repeated measures on an *individual* person, then the covariation represents correlated changes *within* the person. This distinction of *between-* and *within-person* covariance is not trivial. The two levels of analysis are fundamentally different, and only under special conditions can inferences at one level be extended to the other.

People are complex, dynamic biological systems—systems that evolve over time. Nutritional epidemiology is interested in the covariation of certain variables relevant to health. However, the data usually represent the between-person covariation of various factors, yet the inferences are usually intended to be at the within-person level. The implication is that changes in one variable might *lead* to (that is, *cause*) changes in another variable within a person. But the covariance matrix of X, Y, and Z can look very different at the between- and within-person levels. For many biological processes, there is abundant homogeneity in the human population. But individual variation and path dependencies exist as well. Growth, development, and learning are all non-stationary processes (Molenaar 2004).

To the extent that nutrition-related exposures are analyzed to show correlations, the analysis takes place at the between-person level and does not necessarily capture within-person correlations. As one example, Forbes (1984) re-analyzed food diary data and body weight data that seemed perplexing because of an apparent lack of relationship between the two. But, when he plotted within-person change in intake against within-person change in body mass, he observed a nearly perfect correlation.

Although worthwhile, within-person analyses can be more expensive to gather and more complex to analyze. They sometimes necessitate intensive longitudinal measures sufficient to estimate the covariance structure. They also require statistical methods such as time series, vector auto-regressive models, and hidden Markov models, which are not usually taught in many graduate programs.

Additionally, conducting within-person analyses can involve feedback among the X, Y, and Z variables. Should this take place at the within-person level, such feedback can dampen correlations observed at the between-person level.

Although within-person studies may require substantial effort to conduct, it is important to collect the appropriate data to support the intended inferences. To be clear, there *are* many questions that are well-answered with between-person data, such as any health factor where human response is highly uniform. However, these cases are the ones in which the conditions of homogeneity of process are best met. The relevance of between-person covariance to within-person inference is an assumption that must be considered to support more robust research claims.

Incorporating complexities of new measurement methods. Whether OATs or more advanced analyses are used, automated measures such as wearable physical activity monitors can introduce assumptions that also must be considered and corrected for. While the use of newer automation tools in tandem with more traditional observational study methods can help to provide researchers with greater accuracy and clarity, the potential presence of measurement error cannot be discounted.

Nutritional epidemiology lacks uniform approaches to handling such a mountain of exposome data and their relationship to nutrition or health outcomes. Attributing a health risk to a single food or nutrient, as nutritional epidemiology does in studies that often dominate the list of most 'popular' nutrition articles, is no longer entirely defensible given the food-to-exposome, food-to-food, and food-to-other relations among variables. Relying solely on self-reporting as the method of observation is even less defensible, but some of the same or related complexities arise in more objective measurements.

Consider for example, the complex, high-dimensional data collected from physical activity monitors. Ideally, researchers would be able to explore the full algorithms used to generate the end-user data that physical activity monitors provide. But, because manufacturers of commercial-grade wearable technology largely deem these equations proprietary, these algorithms are difficult to obtain except from research-grade devices. Nevertheless, it is possible to account for and resolve measurement error associated with wearable devices, but classic regression methods may be inadequate for this purpose.

Recently, researchers developed and applied novel statistical modeling to correct for wearable device measurement error in a childhood obesity study (Tekwe et al. 2019). Those authors note:

> In this setting, we considered a scalar valued outcome with a functional covariate that was corrupted by measurement error. Most existing methods either implicitly assume the measurement errors are independent over time, or the measurement error covariance is known or can be estimated. However, the measurement errors are likely to be correlated over time. In addition, the measurement error variances are never known and estimates are seldom available. In this paper, we took advantage of the additional information provided in an instrument variable and developed a generalized method of moments-based approach to identify and consistently estimate the functional regression coefficient. (Tekwe et al. 2019)

The researchers illustrated that ignoring measurement error can lead to biased estimations. They add, "We successfully applied our proposed model to conclude that the estimated association between baseline measures of energy expenditure and the 18-month change in BMI was sometimes significant. This association indicated that school programs and policies that increase physical activity among students might have some beneficial impact... Our developed methods improve on the current statistical approaches used to evaluate the effectiveness of such policies" (Tekwe et al. 2019).

Multiple analytical strategies together—including appropriately robust statistical tools—can also mitigate residual confounder, reverse causation, and other biases that commonly plague observational studies (Davey Smith and Ebrahim 2003). Some of these tools have been popularized in other disciplines, and nutritional epidemiology can learn from the groundwork laid by others. We introduce a few such approaches below.

Stronger inferential analyses. Widespread genetic testing in large-scale cohorts promises statistical power sufficient for generating stronger (and often polygenic) analyses for nutritional exposures. These genetically informed analyses are useful for investigating genetic associations and producing more individualized nutrient-outcome predictions. More important to the discussion of causation, Mendelian randomization (MR) using genetic information can be used in these large-scale cohorts to link nutritional exposures to health outcomes.

An adaptation of the instrumental variable approach, MR relies on the genotype as a valid proxy for nutritional (or other types of) exposure and quantifies the causal effect of this proxy on the outcome of interest. Because genotype necessarily precedes any disease outcome, MR conclusions are impervious to reverse causality. Furthermore, because of Mendel's law of independent assortment, both (unlinked) measured and unmeasured confounders are, on average, similarly distributed across genotype/exposure groups, thereby reducing the likelihood of bias due to confounding.

To date, MR has been successfully used to investigate causal effects of exposures related to alcohol and obesity (Au Yeung et al. 2012; Winter-Jensen et al. 2020). In other studies, the causal effect of dairy consumption on a variety of cardiometabolic outcomes was successfully estimated by using lactase persistence polymorphism (LCT-12910C > T) (Mendelian Randomization of Dairy Consumption Working Group 2018; Vissers et al. 2019). LCT-12910C > T has been shown to be a reliable proxy for dairy intake, although its effectiveness as such may vary by population (Chin et al. 2019).

The validity of such MR conclusions generally is predicated on three assumptions. First, a genotype must serve as a strong proxy for the exposure that it is purported to represent. This assumption is often tenuous in nutritional epidemiology. This is especially true for the most controversial exposures such as red meat, eggs, industrially-processed food, or sugar-sweetened beverages. Because of the limited magnitude of genetic effects, often MR studies require very large samples to achieve sufficient statistical power. Yet, even with well-powered studies, confounding by total energy intake remains an almost intractable possibility, threatening the validity of the resulting findings.

The second assumption precludes any horizontally pleiotropic effects of the genotype on the outcome. This condition may be tested using commonly implemented statistical methods and replaced by more lenient assumptions in some MR models (Haycock et al. 2016).

The third assumption excludes any confounding of the relationship between the genetic proxy and the disease outcome and is not directly verifiable. Even with its caveats, MR is another potentially useful analytical tool for nutritional epidemiology. Training in best practices should include selecting appropriate nutritional exposures and their genetic proxies, testing of MR assumptions, choosing appropriate statistical models, and establishing reproducibility of MR findings.

Another approach, used more frequently by econometricians to uncover previously unmeasured biases, uses statistical analyses to create an empirical distribution of non-causal associations. That is, a model is run on the exposure-outcome relationship of interest (e.g., a food's relationship to cardiovascular disease), and on relationships that are not expected to be causally related, which are treated as controls. Finding an association in these presumed non-causal, control relationships may indicate that a common bias explains the association in both the relationship of interest and the control relationships. A related comparison has been used to discuss the causal evidence behind a diet-mortality association (Klurfeld 2015), and investigators on the present paper are involved in using a generalized method sometimes referred to as empirical p-value calibration (Schuemie et al. 2014) to investigate nutrient-disease relationships.

Yet another approach relates to the flexibility in choices in OATs, in which selection biases (intentional or unintentional) in choosing covariates (e.g., age, sex) and operationalization of dietary variables (e.g., dichotomous, continuous) may result in substantially different results. Rather than trying to identify one "best" model, another approach is to test the robustness of the analysis on many different legitimate analytical choices. In the simplest form, this is done in OATs by modeling a bivariate relationship, the selectively adjusted model, and the "kitchen sink" (or inclusion of all covariates) model. A multi-verse of analyses (Steegen et al. 2016), also called vibration of effects (Patel, Burford, and Ioannidis 2015) or specification curve analyses (Simonsohn, Simmons, and Nelson 2019), extends this to test many different model specifications. If the models are not robust to these choices, the nature of any causal relationship between the exposure and outcomes of interest comes into question.

Combining these stronger analytical approaches, along with considering appropriate inference (e.g., between-person *and* within-person designs) and applying novel statistical approaches to datasets, are just some of the many supplemental methods nutritional epidemiology could—and should—be using to move beyond OATs. Doing so is integral to the management and mitigation of alternative explanations and can serve to strengthen nutritional epidemiology's contributions to science.

Stronger execution and reporting

Recommendations

- Nutritional epidemiology should adhere to reporting guidelines (e.g., CONSORT and STROBE-nut).
- To prevent selective non-reporting of studies and results, investigators should register research prospectively (e.g., on ClinicalTrials.gov) and report results for all outcomes and analyses.
- To improve transparency and openness, investigators should share research materials, data, and code.
- To promote scientifically appropriate interpretations, researchers should avoid "spin" in scientific reports and press releases and identify limitations associated with their findings.

Discussion summary

Considering causal inference in nutrition.

What it means to have cause and effect is the same whether an investigator is considering chemical reactions in a tube, pharmaceuticals in people, or social determinants of health. What differs is the ability to probe those questions with gold-standard, causal methodology. The difficulty in answering causal questions in nutrition has resulted in some authors proposing lowering the field's standard of evidence (Katz et al. 2019; Schwingshackl et al. 2016). This includes elevating or disregarding problems with limited-quality assessments, such as FFQs; trusting non-representative, qualitative evidence as causal evidence; and assigning arbitrary point values to (misinterpreted) heterogeneity across studies in nutrition science.

One approach, NutriGrade, suggested down-weighting evidence based on disclosure statements or affiliations (Schwingshackl et al. 2016). However, evaluating evidence based on disclosures is untenable given inherent biases in the field regardless of funding (Ioannidis and Trepanowski 2018). Interpreting science should be limited to the data, the methods, and the logic connecting the data and methods to the results (Brown, Kaiser, and Allison 2018). Indeed, members of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group rebutted incorporation of funding bias in a nutrition-specific grading scheme, stating, "There is no plausible rationale or supporting evidence to justify their approach to include funding bias as a separate item" (Meerpohl et al. 2017). Evidence should be evaluated based on the science, not the scientists.

One challenge in evaluating evidence in a "hierarchy of evidence" is the implicit or explicit assumption that studies are addressing the same question, but often they may not be. Any description of the strength of evidence must include a clear delineation of the research question being asked. In evidence-based medicine, this is often represented by PICOTS elements. These include the Population being studied; the Intervention or exposure thought to cause an effect; the Comparator or Control, which is an alternative to the exposure; the Outcome, which is the health state being assessed; the Time at which the outcome is being assessed, and the Setting or Study design, which includes description of an experimental setting and type.

Frequently, a "hierarchy of evidence" is constructed with meta-analyses of randomized trials perched at the top and with observational evidence, animal studies, and in vitro studies descending the pyramid. Yet, a randomized trial will frequently investigate a well-characterized exposure (the I in PICOTS), such as a defined or specified nutrient or diet, but the trial will evaluate an intermediate outcome, such as blood cholesterol rather than atherosclerosis.

On the other hand, some of the more useful observational studies may use poorly characterized exposures—such as self-reported food frequency estimates extrapolated to actual nutrient quantities—but measure the actual outcome of interest, such as ischemic stroke. In those examples, the randomized controlled trial has a high-quality exposure and inferential design, but it fails to evaluate directly the outcome of interest.

Meanwhile, the observational study has a low-quality exposure and inferential design, but it does study the outcome of interest.

Systems such as GRADE explicitly consider the strength of evidence in causal health claims (Guyatt et al. 2006). Meta-analyses of high-quality randomized controlled trials often lead to the greatest certainty ratings, given their key importance for average causal effects.

Te Morenga and colleagues (2012) communicated the challenge of evaluating nutrition science using approaches like GRADE, by noting that nutrition research may be subject to "potential bias, inconsistency, indirectness, imprecision or reliance on study type other than randomized trials", which results in the downgrading of evidence. They suggested that "formally identifying effects which are regarded as important and based on high quality evidence using the GRADE system may be unattainable in the context of nutritional determinants of chronic disease" (Te Morenga, Mallard, and Mann 2012).

This sentiment was echoed by others in response to the publishing of NutriGrade (a nutrition-specific alternative to—rather than an extension of—GRADE), in which members of the GRADE Working Group remarked that "…lack of blinded randomized controlled trials and the resulting sparse bodies of randomized evidence is not a methodologic shortcoming of the GRADE approach but a limitation of the evidence base" (Meerpohl et al. 2017). The strength of causal evidence is therefore understood to be a property of the science, and that nutritional epidemiology is not in some way exceptional.

In one argument against approaches such as GRADE, the authors of NutriGrade state their group is comprised of nutrition scientists, "whereas GRADE is historically composed of mostly clinical research scientists" and that other disciplines have "found that processing evidence in the clinical research compared with the public health research areas follows slightly different approaches" (Schwingshackl et al. 2017).

This seems to imply that nutrition should not be a clinical science and that public health should be held to lower standards of causal evidence. It is true that the strength of evidence in public health and nutritional epidemiology is, frequently, of lower causal strength compared with some other health-related fields. Rather than adopt low standards of evidence, individual scientists, journals, and scientific societies should embrace transparency and communicate the strengths and limitations of various approaches to nutrition research with greater clarity and nuance.

Example of applying GRADE to nutrition.

GRADE approaches have been successfully used to evaluate and communicate nutrition evidence. NutriRECS (Nutritional Recommendations and accessible Evidence summaries Composed of Systematic reviews) (Johnston et al. 2018), for instance, has evaluated multiple questions in the domain of nutrition using the GRADE approach, including named dietary patterns and weight or cardiovascular disease risk factors; red and processed meat and health outcomes;

probiotics to prevent *Clostridium difficile* infection; and others (Ge et al. 2020; Johnston et al. 2019; Goldenberg, Mertz, and Johnston 2018). The approach demonstrates that guidance consistent with internationally-accepted standards can be achieved in the domain of nutrition.

However, recommendations coming from evidence summaries can sometimes be confusing, even if based on strong methodology at the core. For example, NutriRECS uses the GRADE "evidence to decision" framework (Alonso-Coello, Schünemann, et al. 2016; Alonso-Coello, Oxman, et al. 2016), which includes factors like how much stakeholders value the outcome and whether the intervention would be acceptable. The red and processed meat example mentioned earlier received considerable public attention. Although characterized as a successful use of the approach, the authors establish weak evidence, yet make an active recommendation to *continue* current levels of consumption based on the "evidence to decision" framework. The recommendation process thus risked conflating *conclusions* derived from the science with the *decisions* based on what the target population may prefer. An active recommendation to continue current practice might imply to the audience that changing consumption levels—in either direction—would be deleterious for health, as opposed to being most consistent with preferences.

Nonetheless, much of the public criticism of the NutriRECS conclusion was based on nutrition exceptionalism (i.e., that evidentiary standards should be different in nutrition), or perceived conflicts of interest (i.e., factors unrelated to the data, methods, or conclusions) (Qian et al. 2020; Neuhouser 2020; Leroy and Barnard 2020; Rubin 2020; Oreskes 2021; Vernooij et al. 2021).

Improving research reporting. The strengths of a body of evidence cannot properly be evaluated if the collection of evidence is inadequately reported. All too often nutritional epidemiology is discredited through multiplicity and the selective non-reporting of studies and results. When combined, these may be the greatest contributor to the falsity of scientific claims (Goodman, Fanelli, and Ioannidis 2016).

Studies might evaluate the effect of a nutrient by calculating multiple primary and secondary outcomes. The number of results estimated in a study is a function of both the number of outcome definitions and the number of methods used to analyze those outcomes (Mayo-Wilson, Li, et al. 2017). Even more results are calculated when studies also evaluate the effects of multiple exposures (e.g., nutrients). Because some results will appear to be both clinically and statistically important by chance alone (i.e., false positives), conducting multiple studies and calculating multiple results leads to both true discoveries and false discoveries (Tannock 1996; Greenland 2008). One potential solution is, within a dataset, to make public the number of possible independent variables after accounting for their correlations, and the typical correlations between exposures and outcomes to place new results in their context (Patel and Ioannidis 2014).

Additional methods, as discussed previously, have been developed to assess results from the spectrum of model specifications simultaneously to test the robustness of reasonable analytical choices (Steegen et al. 2016; Simonsohn, Simmons, and Nelson 2020; Patel, Burford, and Ioannidis 2015). Such methods may mitigate bias from the numerous researcher choices during the analysis phase (Gelman and Loken 2013).

Results in journal articles might be systematically biased if they include a disproportionate number of "positive" results, and if the "negative" (e.g., non-significant) results are disproportionately represented in investigators' file drawers (Rosenthal 1979). That notion is supported by direct evidence of study non-publication and by evidence that "primary outcomes" reported in journal articles differ systematically from those reported in study protocols, both of which are related to the significance of results (Chan et al. 2004; Hahn, Williamson, and Hutton 2002; Cooper, DeNeve, and Charlton 1997). Reviews and meta-analyses (Williamson et al. 2005; Williamson and Gamble 2005; Goodman and Dickersin 2011; Mayo-Wilson, Li, et al. 2017) and scientific theories might be incorrect if they depend on a biased subsample of results and *hypothesizing after the results are known* ("HARKing") (Kerr 1998).

The selective non-reporting of studies and results, known as "publication bias" and "outcome reporting bias," respectively, is prevalent in health research (Dwan et al. 2013). Underreporting research has been proposed to be a form of scientific misconduct (Chalmers 1990; Wallach and Krumholz 2019), and some investigators withhold data because of competing interests (Blumenthal et al. 1997). Others fail to submit null findings for publication because they believe their results are uninteresting or unimportant, or that publishers simply will not wish to print them (Chan and Altman 2005; Franco, Malhotra, and Simonovits 2014; Dickersin 1990).

The ability to reproduce results from previous studies is often a hallmark of their truthfulness (Goodman, Fanelli, and Ioannidis 2016). Both multiplicity and selective non-reporting have contributed to irreproducibility in nutritional epidemiology, and the field's dearth of data transparency.

Reproducible workflows and open science. Large-scale statistical modeling, simulation, and data analytics are hindered by a lack of uniformity in software workflows. This has further contributed to the ongoing "reproducibility crisis" in several science domains, including nutritional epidemiology (n.b., we recognize disagreement over calling it a "crisis") (Sweedler 2019). Computational platforms, data sharing frameworks, and archiving of computing environments support reproducibility by lowering barriers to scientific sharing and information preservation (Huo, Nabrzyski, and Vardeman 2015; Open Science Collaboration 2015; Baker 2016). Nutritional epidemiology can take advantage of scientific workflows in order to process large-scale scientific computations in

distributed systems. Workflows and distributed systems have been adopted across scientific domains and have underpinned some of the most significant discoveries of the past several decades (Deelman et al. 2015; Klimentov et al. 2015).

Nutritional epidemiology can also leverage open-source software and open science. The datasets and code used in nutritional epidemiology are rarely made public, hindering reproducibility efforts. Because scientific computing has moved toward the adoption of such tools to perform analyses, data—including input and output datasets, graphs, and intermediate results—are increasingly made available as part of the scientific outcome. Some initiatives have proposed systems, such as RunMyCode.org, Research Compendia, Research Objects, and myExperiment, which facilitate the reproducibility of analyses across processing environments (Goble et al. 2010; Stodden, Hurlin, and Pérignon 2012; Nüst et al. 2017; Bechhofer et al. 2010). Other initiatives have published online, open-source books that share data, code, software versions, or archived computational environments to foster reproducible practices (Kitzes, Turek, and Deniz 2017). These open-source items could then be used as "proof-of-reproducibility" elements in scientific publications or as executable receipts to assist others as they attempt to reproduce equivalent environments.

Registering the details of one's study also would raise the bar for nutritional epidemiology. First proposed for clinical trials (Chalmers and Nadas 1977; Simes 1986; Meinert 1988), study registration is now a widely used method for recording basic details of both trials and observational studies (Nosek et al. 2015), and is a scientific and ethical imperative (World Medical Association 2001; De Angelis et al. 2004). To register a study, investigators enter information about study design and procedures in a public, independently-controlled register. By defining outcomes completely (Zarin et al. 2011; Cybulski, Mayo-Wilson, and Grant 2016) and by registering studies prospectively, sometimes called "preregistration" (Rice and Moher 2019), investigators can improve trust in their findings, link multiple reports about the same study (Mayo-Wilson, Li, et al. 2018), and increase access to their results (Chan et al. 2014). The World Health Organization has defined a minimum dataset and maintains an international list of study registers (De Angelis et al. 2005). The largest register, ClinicalTrials.gov (Zarin et al. 2017), is maintained by the U.S. National Institutes of Health (NIH) and includes both trials and observational studies from around the world (Williams et al. 2010). Because registering and updating registrations requires time and expertise, and because institutions may be ethically and legally responsible for ensuring that studies are registered, universities should support investigators in this process (Mayo-Wilson, Heyward, et al. 2018).

In addition to registers, detailed methods can be published in study protocols (Chan et al. 2013) and statistical analysis plans (SAPs) (Gamble et al. 2017). Protocols and SAPs are useful for minimizing and identifying multiplicity and selective non-reporting. That is, a well-defined outcome can be analyzed using many statistical methods, which will produce different numerical results (Mills 1993; Simmons, Nelson, and Simonsohn 2011). Publishing protocols and statistical analysis plans can help investigators to clarify their hypotheses in advance (Nosek et al. 2018), avoid the temptation to conduct inappropriate analyses (Wang, Yan, and Katz 2018), and identify differences between planned results and the results in their final reports (Pc et al. 2010). These documents are critical components of a system needed to promote rigorous design, measurement, and reporting (Dickersin and Mayo-Wilson 2018).

Developing a core outcome set—the minimum group of outcomes to include in studies of a health condition (Boers et al. 2014)—can also promote consistency across studies and better interpretation of multiple results within studies (e.g., ADOPT standards for obesity) (MacLean et al. 2018). Nutritional epidemiology will benefit when individual researchers willingly commit to a degree of similarity across studies, such as harmonizing experimental definitions and measures of exposures of interest.

Nutritional epidemiology must also communicate the limitations of approaches more clearly and for a broader audience. Traditional news outlets, social media, and others over-interpret weak evidence shared by researchers, journals, and institutions' press offices (Brown, Bohan Brown, and Allison 2013). The spread of conflicting information itself may be problematic by making future communications about nutrition and health difficult to accomplish (Clark, Nagler, and Niederdeppe 2019). Here, too, nutrition is not unique (Selvaraj, Borkar, and Prasad 2014; Haber et al. 2018). Fortunately, a growing number of investigators in nutrition and related fields are not only motivated, but also well-positioned to bring about each of these much-needed reforms.

Discussion

Should such widespread reforms finally begin to take hold, there may be winners and some "losers" in the short term. For instance, academicians making a switch from inexpensive, easily implemented observational study methods to a mix of stronger, more intensive methods may find themselves with fewer opportunities to publish and fewer awarded grants, particularly during the transition. But, over the long term, the field of nutritional epidemiology would have much to gain. Bringing nutritional epidemiology into the realm of more rigorous science would boost the discipline's credibility. It could also help to further strengthen the field by attracting top, new talent. And employing stronger study designs—including a mix of more accurate measures and analyses and more transparent reporting overall—would add value to science as a whole. Armed with more trustworthy results, health care providers and policymakers potentially could make a real and more lasting impact on public health.

Accomplishing such sweeping change will take dedication, time, and patience, but seeking allies with shared interests could help to facilitate nutritional epidemiology's transition. Investigators could take a multi-disciplinary approach, with

nutritional epidemiologists leveraging the expertise of engineers, computational analysts, geneticists, and other outside investigators. Such multi-disciplinary collaboration could engender study designs that are not only more rigorous but also more creative.

Academic journal editors, editorial boards, and peer reviewers can also help drive essential changes to nutritional epidemiological investigation. They can choose to elevate the visibility of more complex studies that make clear contributions to science. They can also enact journal-wide policies requiring study reproducibility, design preregistration, the availability of data repositories, rigorous methods, and more. Both seasoned investigators and those who are new to the field will be incentivized to apply greater scientific rigor to their research efforts. Meanwhile, academicians failing to adapt to journal changes in policy risk being left out of the literature.

By signaling that they, too, demand more scientific rigor, grant-funding agencies could act as drivers of change. Agencies could request specific project types that incorporate the stronger designs, measurements, analyses, and reporting recommendations outlined herein.

Taking on so much change is never easy. But neither is this particular field of study. As author Stuart Ritchie notes:

> Rather like psychology, nutritional epidemiology is *hard.* An incredibly complex physiological and mental machinery is involved in the way we process food and decide what to eat; observational data are subject to enormous noise and the vagaries of human memory; randomised trials can be tripped up by the complexities of their own administration… . Perhaps the very scientific questions that the public wants to have answered the most—what to eat, how to educate children… and so on—are the ones where the science is the murkiest, most difficult, and most self-contradictory. All the more reason that scientists need to take more seriously the task of sensibly communicating their findings to the public. (Ritchie 2020)

And all the more reason to give more than lip service to the need for reform. It is time to *act*—to introduce, teach, promote, and normalize stronger methods of study—and time to elevate nutritional epidemiology to the highest standard.

Acknowledgments

Disclosure statement

ORCID

Andrew W. Brown http://orcid.org/0000-0002-1758-8205
David M. Klurfeld http://orcid.org/0000-0001-7093-3998
Evan Mayo-Wilson http://orcid.org/0000-0001-6126-2459
Patrick D. Quinn http://orcid.org/0000-0002-6770-8762
Carmen Tekwe http://orcid.org/0000-0002-1857-2416
Danny Valdez http://orcid.org/0000-0002-2355-9881
Colby J. Vorland http://orcid.org/0000-0003-4225-372X
David B. Allison http://orcid.org/0000-0003-3566-9399

References

Afshari, R. 2017. Gustav III's risk assessment on coffee consumption; A medical history report. *Avicenna Journal of Phytomedicine* 7 (2):99–100.

Almirall, D., I. Nahum-Shani, N. E. Sherwood, and S. A. Murphy. 2014. Introduction to SMART designs for the development of adaptive interventions: With application to weight loss research. *Translational Behavioral Medicine* 4 (3):260–74. doi: 10.1007/s13142-014-0265-0.

Alonso-Coello, P., A. D. Oxman, J. Moberg, R. Brignardello-Petersen, E. A. Akl, M. Davoli, S. Treweek, R. A. Mustafa, P. O. Vandvik, J. Meerpohl, et al. 2016. GRADE Evidence to Decision (EtD) frameworks: A systematic and transparent approach to making well-informed healthcare choices. 2: Clinical practice guidelines. *BMJ (Clinical Research Ed.)* 353:i2089. doi: 10.1136/bmj.i2089.

Alonso-Coello, P., H. J. Schünemann, J. Moberg, R. Brignardello-Petersen, E. A. Akl, M. Davoli, S. Treweek, R. A. Mustafa, G. Rada, S. Rosenbaum, et al. 2016. GRADE Evidence to Decision (EtD) frameworks: A systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical Research Ed.)* 353:i2016. doi: 10.1136/bmj.i2016.

Au Yeung, S. L., C. Q. Jiang, K. K. Cheng, B. Liu, W. S. Zhang, T. H. Lam, G. M. Leung, and C. M. Schooling. 2012. Evaluation of moderate alcohol use and cognitive function among men using a Mendelian randomization design in the Guangzhou biobank cohort study. *American Journal of Epidemiology* 175 (10):1021–8. doi: 10.1093/aje/kwr462.

Baker, M. 2016. 1,500 Scientists lift the lid on reproducibility. *Nature* 533 (7604):452–4. doi: 10.1038/533452a.

Barron, R., K. Bermingham, L. Brennan, E. R. Gibney, M. J. Gibney, M. F. Ryan, and A. O'Sullivan. 2016. Twin metabolomics: The key to unlocking complex phenotypes in nutrition research. *Nutrition Research (New York, N.Y.)* 36 (4):291–304. doi: 10.1016/j.nutres.2016.01.010.

Basiotis, P. P., S. O. Welsh, F. J. Cronin, J. L. Kelsay, and W. Mertz. 1987. Number of days of food intake records required to estimate individual and group nutrient intakes with defined confidence. *The Journal of Nutrition* 117 (9):1638–41. doi: 10.1093/jn/117.9.1638.

Bechhofer, S., D. De Roure, M. Gamble, C. Goble, and I. Buchan. 2010. Research objects: Towards exchange and reuse of digital knowledge. *Nature Precedings 1-1.* doi: 10.1038/npre.2010.4626.1.

Belluz, J. 2018. This Mediterranean diet study was hugely impactful. The science just fell apart. Accessed December 31, 2020. https://

www.vox.com/science-and-health/2018/6/20/17464906/mediterranean-diet-science-health-predmed.

Blumenthal, D., E. G. Campbell, M. S. Anderson, N. Causino, and K. S. Louis. 1997. Withholding research results in academic life science. Evidence from a national survey of faculty. *JAMA* 277 (15):1224–8.

Boers, M., J. R. Kirwan, G. Wells, D. Beaton, L. Gossec, M.-A. d'Agostino, P. G. Conaghan, C. O. Bingham, P. Brooks, R. Landewé, et al. 2014. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *Journal of Clinical Epidemiology* 67 (7):745–53. doi: 10.1016/j.jclinepi.2013.11.013.

Brown, A. W., M. M. Bohan Brown, and D. B. Allison. 2013. Belief beyond the evidence: Using the proposed effect of breakfast on obesity to show 2 practices that distort scientific evidence. *The American Journal of Clinical Nutrition* 98 (5):1298–308. doi: 10.3945/ajcn.113.064410.

Brown, A. W., J. P. Ioannidis, M. B. Cope, D. M. Bier, and D. B. Allison. 2014. Unscientific beliefs about scientific topics in nutrition. *Advances in Nutrition (Bethesda, MD.)* 5 (5):563–5. doi: 10.3945/an.114.006577.

Brown, A., K. Kaiser, and D. B. Allison. 2018. Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences of the United States of America* 115 (11):2563–70. doi: 10.1073/pnas.1708279115.

Burrows, T. L., Y. Y. Ho, M. E. Rollo, and C. E. Collins. 2019. Validity of dietary assessment methods when compared to the method of doubly labeled water: A systematic review in adults. *Frontiers in Endocrinology* 10:850. doi: 10.3389/fendo.2019.00850.

Cade, J. E., V. J. Burley, D. L. Warm, R. L. Thompson, and B. M. Margetts. 2004. Food-frequency questionnaires: A review of their design, validation and utilisation. *Nutrition Research Reviews* 17 (1):5–22. doi: 10.1079/NRR200370.

Chalmers, I. 1990. Underreporting research is scientific misconduct. *JAMA* 263 (10):1405–8.

Chalmers, T. C., and A. S. Nadas. 1977. Randomize the first patient!. *The New England Journal of Medicine* 296 (2):107. doi: 10.1056/NEJM197701132960214.

Chan, A. W., and D. G. Altman. 2005. Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ (Clinical Research Ed.)* 330 (7494):753. doi: 10.1136/bmj.38356.424606.8F.

Chan, A.-W., A. Hróbjartsson, M. T. Haahr, P. C. Gøtzsche, and D. G. Altman. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* 291 (20):2457–65. doi: 10.1001/jama.291.20.2457.

Chan, A.-W., F. Song, A. Vickers, T. Jefferson, K. Dickersin, P. C. Gøtzsche, H. M. Krumholz, D. Ghersi, and H. B. van der Worp. 2014. Increasing value and reducing waste: Addressing inaccessible research. *The Lancet* 383 (9913):257–66. doi: 10.1016/S0140-6736(13)62296-5.

Chan, A. W., J. M. Tetzlaff, P. C. Gøtzsche, D. G. Altman, H. Mann, J. A. Berlin, K. Dickersin, A. Hróbjartsson, K. F. Schulz, W. R. Parulekar, et al. 2013. SPIRIT 2013 explanation and elaboration: Guidance for protocols of clinical trials. *BMJ (Clinical Research Ed.)* 346:e7586. doi: 10.1136/bmj.e7586.

Chin, E. L., L. Huang, Y. Y. Bouzid, C. P. Kirschke, B. Durbin-Johnson, L. M. Baldiviez, E. L. Bonnel, N. L. Keim, I. Korf, C. B. Stephensen, et al. 2019. Association of lactase persistence genotypes (rs4988235) and ethnicity with dairy intake in a healthy U.S. population. *Nutrients* 11 (8):1860. doi: 10.3390/nu11081860.

Chiu, Y.-H., J. E. Chavarro, B. A. Dickerman, J. E. Manson, K. J. Mukamal, K. M. Rexrode, E. B. Rimm, and M. A. Hernán. 2021. Estimating the effect of nutritional interventions using observational data: The American Heart Association's 2020 Dietary Goals and mortality. *American Journal of Clinical Nutrition* 114 (2):690–703. doi: 10.1093/ajcn/nqab100.

Clark, D., R. H. Nagler, and J. Niederdeppe. 2019. Confusion and nutritional backlash from news media exposure to contradictory information about carbohydrates and dietary fats. *Public Health Nutrition* 22 (18):3336–48. doi: 10.1017/S1368980019002866.

Cofield, S. S., R. V. Corona, and, and D. B. Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity Facts* 3 (6):353–6. doi: 10.1159/000322940.

Colen, C. G., and D. M. Ramey. 2014. Is breast truly best? Estimating the effects of breastfeeding on long-term child health and wellbeing in the United States using sibling comparisons. *Social Science & Medicine (1982)* 109:55–65. doi: 10.1016/j.socscimed.2014.01.027.

Collins, L. M. 2018. *Optimization of behavioral, biobehavioral, and biomedical interventions: The multiphase optimization strategy (MOST).* Cham, Switzerland: Springer International Publishing.

Cooper, H., K. DeNeve, and K. Charlton. 1997. Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods* 2 (4):447–52. doi: 10.1037/1082-989X.2.4.447.

Cybulski, L., E. Mayo-Wilson, and S. Grant. 2016. Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *The Journal of Consulting and Clinical Psychology* 84 (9):753–67. doi: 10.1037/ccp0000115.

D'Onofrio, B. M., Q. A. Class, M. E. Rickert, A. C. Sujan, H. Larsson, R. Kuja-Halkola, A. Sjölander, C. Almqvist, P. Lichtenstein, A. S. Oberg, et al. 2016. Translational epidemiologic approaches to understanding the consequences of early-life exposures. *Behavior Genetics* 46 (3):315–28. doi: 10.1007/s10519-015-9769-8.

Davey Smith, G., and S. Ebrahim. 2003. Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* 32 (1):1–22. doi: 10.1093/ije/dyg070.

De Angelis, C. D., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. M. Overbeke, et al. 2005. Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors *Annals of Internal Medicine* 143 (2):146–8. doi: 10.7326/0003-4819-143-2-200507190-00016.

De Angelis, C., J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, A. J. P. M. Overbeke, et al. 2004. Clinical trial registration: A statement from the international committee of medical journal editors. *The New England Journal of Medicine* 351 (12):1250–1. doi: 10.1056/NEJMe048225.

Deelman, E., K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, et al. 2015. Pegasus, a workflow management system for science automation. *Future Generation Computer Systems* 46:17–35. doi: 10.1016/j.future.2014.10.008.

Dhurandhar, N. V., D. Schoeller, A. W. Brown, S. B. Heymsfield, D. Thomas, T. I. Sørensen, J. R. Speakman, M. Jeansonne, and D. B. Allison. 2015. Energy balance measurement: When something is not better than nothing. *International Journal of Obesity (2005)* 39 (7):1109–13. doi: 10.1038/ijo.2014.199.

Dickersin, K. 1990. The existence of publication bias and risk factors for its occurrence. *JAMA* 263 (10):1385–9. doi: 10.1001/jama.1990.03440100097014.

Dickersin, K., and E. Mayo-Wilson. 2018. Standards for design and measurement would make clinical research reproducible and usable. *Proceedings of the National Academy of Sciences of the United States of America* 115 (11):2590–4. doi: 10.1073/pnas.1708273114.

Duan, N., R. L. Kravitz, and C. H. Schmid. 2013. Single-patient (n-of-1) trials: A pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology* 66 (8 Suppl):S21–S8.

Dwan, K., C. Gamble, P. R. Williamson, and J. J. Kirkham. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - An updated review. *PLoS One* 8 (7):e66844. doi: 10.1371/journal.pone.0066844.

Ejima, K., A. W. Brown, D. A. Schoeller, S. B. Heymsfield, E. J. Nelson, and D. B. Allison. 2019. Does exclusion of extreme reporters of energy intake (the "Goldberg cutoffs") reliably reduce or eliminate bias in nutrition studies? Analysis with illustrative associations of energy intake with health outcomes. *American Journal of Clinical Nutrition* 110 (5):1231–9. doi: 10.1093/ajcn/nqz198.

Ferdinand, O. A., B. Sen, S. Rahurkar, S. Engler, and N. Menachemi. 2012. The relationship between built environments and physical activity: A systematic review. *American Journal of Public Health* 102 (10):e7–13.

Fontana, J. M., J. A. Higgins, S. C. Schuckers, F. Bellisle, Z. Pan, E. L. Melanson, M. R. Neuman, and E. Sazonov. 2015. Energy intake estimation from counts of chews and swallows. *Appetite* 85:14–21. doi: 10.1016/j.appet.2014.11.003.

Forbes, G. B. 1984. Energy intake and body weight: A reexamination of two "classic" studies. *The American Journal of Clinical Nutrition* 39 (3):349–50. doi: 10.1093/ajcn/39.3.349.

Franco, A., N. Malhotra, and G. Simonovits. 2014. Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science (New York, N.Y.)* 345 (6203):1502–5. doi: 10.1126/science.1255484.

Friedman, L. M., C. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. 2010. *Fundamentals of clinical trials*. New York: Springer.

Frisell, T., S. Öberg, R. Kuja-Halkola, and A. Sjölander. 2012. Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology (Cambridge, Mass.)* 23 (5):713–20. doi: 10.1097/EDE.0b013e31825fa230.

Gamble, C., A. Krishan, D. Stocken, S. Lewis, E. Juszczak, C. Doré, P. R. Williamson, D. G. Altman, A. Montgomery, P. Lim, et al. 2017. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA* 318 (23):2337–43. doi: 10.1001/jama.2017.18556.

Ge, L., B. Sadeghirad, G. D. C. Ball, B. R. da Costa, C. L. Hitchcock, A. Svendrovski, R. Kiflen, K. Quadri, H. Y. Kwon, M. Karamouzian, et al. 2020. Comparison of dietary macronutrient patterns of 14 popular named dietary programmes for weight and cardiovascular risk factor reduction in adults: Systematic review and network meta-analysis of randomised trials. *BMJ* 369:m696. doi: 10.1136/bmj.m696.

Gelman, A., and E. Loken. 2001. 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. Last Modified November 14, 2013. Accessed October 6, 2021. http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf.

George, B. J., P. Li, H. R. Lieberman, G. Pavela, A. W. Brown, K. R. Fontaine, M. M. Jeansonne, G. R. Dutton, A. J. Idigo, M. A. Parman, et al. 2018. Randomization to randomization probability: Estimating treatment effects under actual conditions of use. *Psychological Methods* 23 (2):337–50. doi: 10.1037/met0000138.

Goble, C. A., J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, S. Bechhofer, M. Roos, P. Li, et al. 2010. myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research* 38 (Web Server issue):W677–82. doi: 10.1093/nar/gkq429.

Goldenberg, J. Z., D. Mertz, and B. C. Johnston. 2018. Probiotics to prevent Clostridium difficile infection in patients receiving antibiotics. *JAMA* 320 (5):499–500. doi: 10.1001/jama.2018.9064.

Goldsby, T., B. George, V. Yeager, B. Sen, A. Ferdinand, D. Sims, B. Manzella, A. Cockrell Skinner, D. Allison, and N. Menachemi. 2016. Urban park development and pediatric obesity rates: A quasi-experiment using electronic health record data. *International Journal of Environmental Research and Public Health* 13 (4):411. doi: 10.3390/ijerph13040411.

Goodman, S., and K. Dickersin. 2011. Metabias: A challenge for comparative effectiveness research. *Annals of Internal Medicine* 155 (1):61–2. doi: 10.7326/0003-4819-155-1-201107050-00010.

Goodman, S., D. Fanelli, and J. P. Ioannidis. 2016. What does research reproducibility mean? *Science Translational Medicine* 8 (341):341ps12. doi: 10.1126/scitranslmed.aaf5027.

Granic, A., R. Andel, A. K. Dahl, M. Gatz, and N. L. Pedersen. 2013. Midlife dietary patterns and mortality in the population-based study of Swedish twins. *Journal of Epidemiology and Community Health* 67 (7):578–86. doi: 10.1136/jech-2012-201780.

Greenland, S. 2008. Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* 167 (5):523–9. doi: 10.1093/aje/kwm355.

Guyatt, G., G. Vist, Y. Falck-Ytter, R. Kunz, N. Magrini, and H. Schunemann. 2006. An emerging consensus on grading recommendations? *Evidence-Based Medicine* 11 (1):2–4. doi: 10.1136/ebm.11.1.2-a.

Haber, N., E. R. Smith, E. Moscoe, K. Andrews, R. Audy, W. Bell, A. T. Brennan, A. Breskin, J. C. Kane, M. Karra, et al. 2018. Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. *PloS One* 13 (5):e0196346. doi: 10.1371/journal.pone.0196346.

Hahn, S., P. R. Williamson, and J. L. Hutton. 2002. Investigation of within-study selective reporting in clinical research: Follow-up of applications submitted to a local research ethics committee. *Journal of Evaluation in Clinical Practice* 8 (3):353–9. doi: 10.1046/j.1365-2753.2002.00314.x.

Haycock, P. C., S. Burgess, K. H. Wade, J. Bowden, C. Relton, and G. Davey Smith. 2016. Best (but oft-forgotten) practices: The design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition* 103 (4):965–78. doi: 10.3945/ajcn.115.118216.

Hayes, M. J., V. Kaestner, S. Mailankody, and V. Prasad. 2018. Most medical practices are not parachutes: A citation analysis of practices felt by biomedical authors to be analogous to parachutes. *CMAJ Open* 6 (1):E31–8. doi: 10.9778/cmajo.20170088.

Hébert, J. R., E. A. Frongillo, S. A. Adams, G. M. Turner-McGrievy, T. G. Hurley, D. R. Miller, and I. S. Ockene. 2016. Perspective: Randomized controlled trials are not a panacea for diet-related research. *Advances in Nutrition (Bethesda, MD.)* 7 (3):423–32. doi: 10.3945/an.115.011023.

Hemming, K., T. P. Haines, P. J. Chilton, A. J. Girling, and R. J. Lilford. 2015. The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ (Clinical Research Ed.)* 350:h391. doi: 10.1136/bmj.h391.

Hernán, M. A. 2018. The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health* 108 (5):616–9. doi: 10.2105/AJPH.2018.304337.

Hill, A. B. 1965. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine* 58 (5):295–300.

Hoover, A., and E. Sazonov. 2016. Measuring human energy intake and ingestive behavior: Challenges and opportunities. *IEEE Pulse.* 7 (6):6–7. doi: 10.1109/MPUL.2016.2606465.

Huo, D., J. Nabrzyski, and C. F. VardemanII. 2015. An ontology design pattern towards preservation of computational experiments. Bethlehem, PA: Proceedings of the 5th Workshop on Linked Science.

Imbens, G. W., and D. B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge, UK: Cambridge University Press.

Ioannidis, J. P. A., and J. F. Trepanowski. 2018. Disclosures in nutrition research: Why it is different. *JAMA* 319 (6):547–8. doi: 10.1001/jama.2017.18571.

Jepsen, P., S. P. Johnsen, M. W. Gillman, and H. T. Sørensen. 2004. Interpretation of observational studies. *Heart (British Cardiac Society)* 90 (8):956–60. doi: 10.1136/hrt.2003.017269.

Johnston, B. C., P. Alonso-Coello, M. M. Bala, D. Zeraatkar, M. Rabassa, C. Valli, C. Marshall, R. El Dib, R. W. Vernooij, P. O. Vandvik, et al. 2018. Methods for trustworthy nutritional recommendations NutriRECS (Nutritional Recommendations and accessible Evidence summaries Composed of Systematic reviews): A protocol. *BMC Medical Research Methodology* 18 (1):1–11. doi: 10.1186/s12874-018-0621-8.

Johnston, B. C., D. Zeraatkar, M. A. Han, R. W. M. Vernooij, C. Valli, R. El Dib, C. Marshall, P. J. Stover, S. Fairweather-Taitt, G. Wójcik, et al. 2019. Unprocessed red meat and processed meat consumption: Dietary guideline recommendations from the Nutritional Recommendations (NutriRECS) Consortium. *Annals of Internal Medicine* 171 (10):756–64. doi: 10.7326/M19-1621.

Kaprio, J., and M. Koskenvuo. 1989. Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs. *Social Science & Medicine (1982)* 29 (9):1083–9. doi: 10.1016/0277-9536(89)90020-8.

Katz, D. L. 2019. Science advisory: Children should keep running with scissors! (except, of course, not really) Accessed January 13, 2021.

https://www.linkedin.com/pulse/science-advisory-children-should-keep-running-except-david/.

Katz, D. L., M. C. Karlsen, M. Chung, M. M. Shams-White, L. W. Green, J. Fielding, A. Saito, and W. Willett. 2019. Hierarchies of evidence applied to lifestyle Medicine (HEALM): Introduction of a strength-of-evidence approach based on a methodological systematic review. *BMC Medical Research Methodology* 19 (1):178. doi: 10.1186/s12874-019-0811-z.

Kerr, N. L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc* 2 (3):196–217. doi: 10.1207/s15327957pspr0203_4.

Kim, S. Y., J. Flory, and C. Relton. 2018. Ethics and practice of trials within cohorts: An emerging pragmatic trial design. *Clinical Trials (London, England)* 15 (1):9–16. doi: 10.1177/1740774517746620.

Kipnis, V., D. Midthune, L. Freedman, S. Bingham, N. E. Day, E. Riboli, P. Ferrari, and R. J. Carroll. 2002. Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition* 5 (6A):915–23. doi: 10.1079/PHN2002383.

Kitzes, J., D. Turek, and F. Deniz, eds. 2017. *The practice of reproducible research: Case studies and lessons from the data-intensive sciences.* Berkeley, CA: University of California Press.

Klimentov, A., P. Buncic, K. De, S. Jha, T. Maeno, R. Mount, P. Nilsson, D. Oleynik, S. Panitkin, A. Petrosyan, et al. 2015. Next generation workload management system for big data on heterogeneous distributed computing. *Journal of Physics: Conference Series* 608 (1):012040. doi: 10.1088/1742-6596/608/1/012040.

Klurfeld, D. M. 2015. Research gaps in evaluating the relationship of meat and health. *Meat Science* 109:86–95. doi: 10.1016/j.meatsci.2015.05.022.

Krall, E. A., and J. T. Dwyer. 1987. Validity of a food frequency questionnaire and a food diary in a short-term recall situation. *Journal of the American Dietetic Association* 87 (10):1374–7. doi: 10.1016/S0002-8223(21)03325-3.

Lear, S. 2019. Should you eat red meat? Navigating a world of contradicting studies. Accessed December 31, 2020. https://www.discovermagazine.com/health/should-you-eat-red-meat-navigating-a-world-of-contradicting-studies#.XaM8HPlKiMo.

Lei, H., I. Nahum-Shani, K. Lynch, D. Oslin, and S. A. Murphy. 2012. A "SMART" design for building individualized treatment sequences. *Annual Review of Clinical Psychology* 8:21–48.

Leigh, A. 2018. *Randomistas: How radical researchers are changing our world.* New Haven, CT: Yale University Press.

Leroy, F., and N. D. Barnard. 2020. Children and adults should avoid consuming animal products to reduce risk for chronic disease: NO. *American Journal of Clinical Nutrition* 112 (4):931–6. doi: 10.1093/ajcn/nqaa236.

Lichtenstein, P., U. De Faire, B. Floderus, M. Svartengren, P. Svedberg, and N. L. Pedersen. 2002. The Swedish Twin Registry: A unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine* 252 (3):184–205. doi: 10.1046/j.1365-2796.2002.01032.x.

MacLean, P. S., Rothman, A. J. Nicastro, H. L. Czajkowski, S. M. Agurs-Collins, T. Rice, E. L. Courcoulas, A. P. Ryan, D. H. Bessesen, and D. H. Loria. 2018. The accumulating data to optimally predict obesity treatment (ADOPT) core measures project: Rationale and approach. *Obesity* 26:S6–S15. doi: 10.1002/oby.22154.

Mayo-Wilson, E., N. Fusco, T. Li, H. Hong, J. K. Canner, and K. Dickersin. 2017. Multiple outcomes and analyses in clinical trials create challenges for interpretation and research synthesis. *Journal of Clinical Epidemiology* 86:39–50. doi: 10.1016/j.jclinepi.2017.05.007.

Mayo-Wilson, E., J. Heyward, A. Keyes, J. Reynolds, S. White, N. Atri, G. C. Alexander, A. Omar, and D. E. Ford. 2018. Clinical trial registration and reporting: A survey of academic organizations in the United States. *BMC Medicine* 16 (1):60. doi: 10.1186/s12916-018-1042-6.

Mayo-Wilson, E., T. Li, N. Fusco, L. Bertizzolo, J. K. Canner, T. Cowley, P. Doshi, J. Ehmsen, G. Gresham, N. Guo, et al. 2017. Cherry-picking by trialists and meta-analysts can drive conclusions about intervention efficacy. *Journal of Clinical Epidemiology* 91:95–110. doi: 10.1016/j.jclinepi.2017.07.014.

Mayo-Wilson, E., T. Li, N. Fusco, and K. Dickersin. 2018. Practical guidance for using multiple data sources in systematic reviews and meta-analyses (with examples from the MUDS study). *Research Synthesis Methods* 9 (1):2–12. doi: 10.1002/jrsm.1277.

McGue, M., M. Osler, and K. Christensen. 2010. Causal inference and observational research: The utility of twins. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 5 (5):546–56. doi: 10.1177/1745691610383511.

Meerpohl, J. J., C. E. Naude, P. Garner, R. A. Mustafa, and H. J. Schünemann. 2017. Comment on "Perspective: NutriGrade: A scoring system to assess and judge the meta-evidence of randomized controlled trials and cohort studies in nutrition Research." *Advances in Nutrition (Bethesda, MD.)* 8 (5):789–90. doi: 10.3945/an.117.016188.

Meinert, C. 1986. *Clinical trials - Design, conduct, and analysis.* New York: Oxford University Press.

Meinert, C. L. 1988. Toward prospective registration of clinical trials. *Controlled Clinical Trials* 9 (1):1–5. doi: 10.1016/0197-2456(88)90002-5.

Mendelian Randomization of Dairy Consumption Working Group. 2018. Dairy consumption and body mass index among adults: Mendelian randomization analysis of 184802 individuals from 25 studies. *Clinical Chemistry* 64 (1):183–91.

Metzger, M. W., and T. W. McDade. 2010. Breastfeeding as obesity prevention in the United States: A sibling difference model. *American Journal of Human Biology: The Official Journal of the Human Biology Council* 22 (3):291–6. doi: 10.1002/ajhb.20982.

Mills, J. L. 1993. Data torturing. *The New England Journal of Medicine* 329 (16):1196–9. doi: 10.1056/NEJM199310143291613.

Molenaar, P. C. 2004. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement* 2 (4):201–18.

Morgan, S. L., and C. Winship. 2015. *Counterfactuals and causal inference.* Cambridge, UK: Cambridge University Press.

Moshfegh, A. J., D. G. Rhodes, D. J. Baer, T. Murayi, J. C. Clemens, W. V. Rumpler, D. R. Paul, R. S. Sebastian, K. J. Kuczynski, L. A. Ingwersen, et al. 2008. The US Department of Agriculture automated multiple-pass method reduces bias in the collection of energy intakes. *The American Journal of Clinical Nutrition* 88 (2):324–32. doi: 10.1093/ajcn/88.2.324.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and replicability in science.* Washington, DC: National Academies Press.

Naukkarinen, J., A. Rissanen, J. Kaprio, and K. H. Pietiläinen. 2012. Causes and consequences of obesity: The contribution of recent twin studies. *International Journal of Obesity (2005)* 36 (8):1017–24. doi: 10.1038/ijo.2011.192.

Neuhouser, M. L. 2020. Red and processed meat: More with less? *The American Journal of Clinical Nutrition* 111 (2):252–5. doi: 10.1093/ajcn/nqz294.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. 2015. Scientific standards. Promoting an open research culture. *Science (New York, N.Y.)* 348 (6242):1422–5. doi: 10.1126/science.aab2374.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America* 115 (11):2600–6. doi: 10.1073/pnas.1708274114.

Nüst, D., M. Konkol, E. Pebesma, C. Kray, M. Schutzeichel, H. Przibytzin, and J. Lorenz. 2017. Opening the publication process with executable research compendia. *D-Lib Magazine* 23 (1/2):10–1045. doi: 10.1045/january2017-nuest.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349 (6251):aac4716. doi: 10.1126/science.aac4716

Oreskes, N. 2021. So is it okay to eat more red and processed meat? Accessed January 13, 2021. https://www.scientificamerican.com/article/so-is-it-okay-to-eat-more-red-and-processed-meat/.

Pallister, T., T. D. Spector, and C. Menni. 2014. Twin studies advance the understanding of gene-environment interplay in human nutrigenomics. *Nutrition Research Reviews* 27 (2):242–51. doi: 10.1017/S095442241400016X.

Patel, C. J., and J. P. Ioannidis. 2014. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *Journal of Epidemiology and Community Health* 68 (11):1096–100. doi: 10.1136/jech-2014-204195.

Patel, C. J., B. Burford, and J. P. Ioannidis. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* 68 (9):1046–58. doi: 10.1016/j.jclinepi.2015.05.029.

Pavela, G., H. Wiener, K. R. Fontaine, D. A. Fields, J. D. Voss, and D. B. Allison. 2015. Packet randomized experiments for eliminating classes of confounders. *European Journal of Clinical Investigation* 45 (1):45–55. doi: 10.1111/eci.12378.

Pc, K. J. K., A. Hróbjartsson, H. Mann, K. Dickersin, J. Berlin, C. Doré, and H. Sox. 2010. A catalogue of reporting guidelines for health research. *BMC Medicine* 8 (18):20334633.

Prentice, A. M., A. E. Black, W. A. Coward, H. L. Davies, G. R. Goldberg, P. R. Murgatroyd, J. Ashford, M. Sawyer, and R. G. Whitehead. 1986. High levels of energy expenditure in obese women. *British Medical Journal (Clinical Research Ed.)* 292 (6526):983–7. doi: 10.1136/bmj.292.6526.983.

Qian, F., M. C. Riddle, J. Wylie-Rosett, and F. B. Hu. 2020. Red and processed meats and health risks: How strong is the evidence? *Diabetes Care* 43 (2):265–71. doi: 10.2337/dci19-0063.

Rhodes, D. G., T. Murayi, J. C. Clemens, D. J. Baer, R. S. Sebastian, and A. J. Moshfegh. 2013. The USDA automated multiple-pass method accurately assesses population sodium intakes. *The American Journal of Clinical Nutrition* 97 (5):958–64. doi: 10.3945/ajcn.112.044982.

Rice, D. B., and D. Moher. 2019. Curtailing the use of preregistration: A misused term. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 14 (6):1105–8. doi: 10.1177/1745691619858427.

Richardson, M. B., M. S. Williams, K. R. Fontaine, and D. B. Allison. 2017. The development of scientific evidence for health policies for obesity: Why and how? *International Journal of Obesity (2005)* 41 (6):840–8. doi: 10.1038/ijo.2017.71.

Ritchie, S. 2020. *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth.* New York, NY: Metropolitan Books.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1):41–55. doi: 10.1093/biomet/70.1.41.

Rosenthal, R. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86 (3):638–41. doi: 10.1037/0033-2909.86.3.638.

Rubin, R. 2020. Backlash over meat dietary recommendations raises questions about corporate ties to nutrition scientists. *JAMA* 323 (5):401–4. doi: 10.1001/jama.2019.21441.

Sacristán, J. A., and T. Dilla. 2018. Pragmatic trials revisited: Applicability is about individualization. *Journal of Clinical Epidemiology* 99:164–6. doi: 10.1016/j.jclinepi.2018.02.003.

Salley, J., E. Muth, M. Wilson, and A. Hoover. 2016. A comparison between human and bite-based methods of estimating caloric intake. *Journal of the Academy of Nutrition and Dietetics* 116 (10):1568–77. doi: 10.1016/j.jand.2016.03.007.

Satija, A., E. Yu, W. C. Willett, and F. B. Hu. 2015. Understanding nutritional epidemiology and its role in policy. *Advances in Nutrition (Bethesda, MD.)* 6 (1):5–18. doi: 10.3945/an.114.007492.

Schoeller, D.A., and M. Westerterp, eds. 2017. *Advances in the assessment of dietary intake.* Boca Raton, FL: CRC Press.

Schuemie, M. J., P. B. Ryan, W. DuMouchel, M. A. Suchard, and D. Madigan. 2014. Interpreting observational studies: Why empirical calibration is needed to correct p-values. *Statistics in Medicine* 33 (2):209–18. doi: 10.1002/sim.5925.

Schwingshackl, L., S. Knüppel, C. Schwedhelm, G. Hoffmann, B. Missbach, M. Stelmach-Mardas, S. Dietrich, F. Eichelmann, E. Kontopantelis, K. Iqbal, et al. 2016. Perspective: NutriGrade: A scoring system to assess and judge the meta-evidence of randomized controlled trials and cohort studies in nutrition research. *Advances in Nutrition (Bethesda, MD.)* 7 (6):994–1004. doi: 10.3945/an.116.013052.

Schwingshackl, L., S. Knüppel, C. Schwedhelm, G. Hoffmann, B. Missbach, M. Stelmach-Mardas, S. Dietrich, F. Eichelmann, E. Kontopantelis, K. Iqbal, et al. 2017. Reply to JJ Meerpohl et al. *Advances in Nutrition (Bethesda, MD.)* 8 (5):790–1. doi: 10.3945/an.117.016469.

Scisco, J., E. Muth, and A. Hoover. 2014. Examining the utility of a bite-count based measure of eating activity in free-living human beings. *Journal of the Academy of Nutrition and Dietetics* 114 (3):464–9. doi: 10.1016/j.jand.2013.09.017.

Selvaraj, S., D. S. Borkar, and V. Prasad. 2014. Media coverage of medical journals: Do the best articles make the news? *PLoS One* 9 (1):e85355. doi: 10.1371/journal.pone.0085355.

Shadish, W. R., D. T. Campbell, and T. D. Cook. 2001. *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Simes, R. J. 1986. Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 4 (10):1529–41. doi: 10.1200/JCO.1986.4.10.1529.

Simmons, J. P., L. D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 (11):1359–66. doi: 10.1177/0956797611417632.

Simonsohn, U., J. P. Simmons, and L. D. Nelson. 2019. Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available at SSRN 2694998. doi: 10.2139/ssrn.2694998.

Simonsohn, U., J. P. Simmons, and L. D. Nelson. 2020. Specification curve analysis. *Nature Human Behaviour* 4 (11):1208–14. doi: 10.1038/s41562-020-0912-z.

Speakman, J. R., Y. Yamada, H. Sagayama, E. S. F. Berman, P. N. Ainslie, L. F. Andersen, L. J. Anderson, L. Arab, I. Baddou, K. Bedu-Addo, et al. 2021. A standard calculation methodology for human doubly labeled water studies. *Cell Reports Medicine* 2 (2):100203. doi: 10.1016/j.xcrm.2021.100203.

Steegen, S., F. Tuerlinckx, A. Gelman, and W. Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 11 (5):702–12. doi: 10.1177/1745691616658637.

Stodden, V., C. Hurlin, and C. Pérignon. 2012. RunMyCode.org: A novel dissemination and collaboration platform for executing published computational results, 1–8. Chicago, IL: 2012 IEEE 8th International Conference on E-Science.

Subar, A. F., S. I. Kirkpatrick, B. Mittl, T. P. Zimmerman, F. E. Thompson, C. Bingley, G. Willis, N. G. Islam, T. Baranowski, S. McNutt, et al. 2012. The Automated Self-Administered 24-hour dietary recall (ASA24): A resource for researchers, clinicians, and educators from the National Cancer Institute. *Journal of the Academy of Nutrition and Dietetics* 112 (8):1134–7. doi: 10.1016/j.jand.2012.04.016.

Sweedler, J. V. 2019. Reproducibility and replicability. *Analytical Chemistry* 91 (13):7971–2. doi: 10.1021/acs.analchem.9b02719.

Tannock, I. F. 1996. False-positive results in clinical trials: Multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute* 88 (3–4):206–7. doi: 10.1093/jnci/88.3-4.206.

Te Morenga, L., S. Mallard, and J. Mann. 2012. Dietary sugars and body weight: Systematic review and meta-analyses of randomised controlled trials and cohort studies. *BMJ* 346 (3):e7492–e7492. doi: 10.1136/bmj.e7492.

Teicholz, N., and G. Taubes. 2018. Rapid response to: Dietary guidelines and health—Is nutrition science up to the task? Accessed December 31, 2020. https://www.bmj.com/content/360/bmj.k822/rr-13.

Tekwe, C. D., R. S. Zoh, M. Yang, R. J. Carroll, G. Honvoh, D. B. Allison, M. Benden, and L. Xue. 2019. Instrumental variable approach to estimating the scalar-on-function regression model with measurement error with application to energy expenditure assessment in childhood obesity. *Statistics in Medicine* 38 (20):3764–81. doi: 10.1002/sim.8179.

Tobias, D. K., and M. Lajous. 2021. What would the trial be? Emulating randomized dietary intervention trials to estimate causal effects with observational data. *The American Journal of Clinical Nutrition* 114 (2):416–7. doi: 10.1093/ajcn/nqab169.

Trepanowski, J. F., and J. P. A. Ioannidis, 2018. Perspective: Limiting dependence on nonrandomized studies and improving randomized trials in human nutrition research: Why and how. *Advances in Nutrition (Bethesda, MD.)* 9 (4):367–77. doi: 10.1093/advances/nmy014.

Vernooij, R. W. M., G. H. Guyatt, D. Zeraatkar, M. A. Han, C. Valli, R. El Dib, P. Alonso-Coello, M. M. Bala, and B. C. Johnston, 2021. Reconciling contrasting guideline recommendations on red and processed meat for health outcomes. *Journal of Clinical Epidemiology* S0895–4356 (21): 00216-X. doi: 10.1016/j.jclinepi.2021.07.008 Online ahead of print

Vissers, L. E. T., I. Sluijs, Y. T. van der Schouw, N. G. Forouhi, F. Imamura, S. Burgess, A. Barricarte, H. Boeing, C. Bonet, M.-D. Chirlaque, et al. 2019. Dairy product intake and risk of type 2 diabetes in EPIC-InterAct: A Mendelian randomization study. *Diabetes Care* 42 (4):568–75. doi: 10.2337/dc18-2034.

Walker, J. L., S. Ardouin, and T. Burrows, 2018. The validity of dietary assessment methods to accurately measure energy intake in children and adolescents who are overweight or obese: A systematic review. *European Journal of Clinical Nutrition* 72 (2):185–97. doi: 10.1038/s41430-017-0029-2.

Wallach, J. D., and H. M. Krumholz, 2019. Not reporting results of a clinical trial is academic misconduct. *Annals of Internal Medicine* 171 (4):293–4. doi: 10.7326/M19-1273. [published Online First: 2019/05/07]

Wang, M. Q., A. F. Yan, and R. V. Katz, 2018. Researcher requests for inappropriate analysis and reporting: A US survey of consulting biostatisticians. *Annals of Internal Medicine* 169 (8):554–8. doi: 10.7326/M18-1230. [published Online First: 2018/10/12]

Weathers, D., J. Siemens, and S. Kopp, 2017. Tracking food intake as bites: Effects on cognitive resources, eating enjoyment, and self-control. *Appetite* 111:23–31. doi: 10.1016/j.appet.2016.12.018.

Westfall, J., and T. Yarkoni, 2016. Statistically controlling for confounding constructs is harder than you think. *PLoS One* 11 (3):e0152719. doi: 10.1371/journal.pone.0152719.

Williams, R. J., T. Tse, W. R. Harlan, and D. A. Zarin. 2010. Registration of observational studies: Is it time? *CMAJ: Canadian Medical Association Journal = Journal de L'Association Medicale Canadienne* 182 (15):1638–42. doi: 10.1503/cmaj092252.

Williamson, P. R., and C. Gamble, 2005. Identification and impact of outcome selection bias in meta-analysis. *Statistics in Medicine* 24 (10):1547–61. doi: 10.1002/sim.2025.

Williamson, P. R., C. Gamble, D. G. Altman, and J. L. Hutton, 2005. Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research* 14 (5):515–24. doi: 10.1191/0962280205sm415oa. [published Online First: 2005/10/27]

Willett, W. C., R. D. Reynolds, S. Cottrell-Hoehner, L. Sampson, and M. L. Browne, 1987. Validation of a semi-quantitative food frequency questionnaire: Comparison with a 1-year diet record. *Journal of the American Dietetic Association* 87 (1):43–7. doi: 10.1016/S0002-8223(21)03057-1.

Windsor, R., T. Baranowski, G. Cutter, and R. A. Windsor. 2001. *Evaluation of health promotion, health education, and disease prevention programs with PowerWeb*. New York, NY: McGraw-Hill Higher Education.

Winter-Jensen, M., S. Afzal, T. Jess, B. G. Nordestgaard, and K. H. Allin, 2020. Body mass index and risk of infections: A Mendelian randomization study of 101,447 individuals. *European Journal of Epidemiology* 35 (4):347–54. doi: 10.1007/s10654-020-00630-7.

World Medical Association. 2001. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization* 79 (4):373–4.

Zarin, D. A., T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide, 2011. The ClinicalTrials.gov results database-update and key issues. *The New England Journal of Medicine* 364 (9):852–60. doi: 10.1056/NEJMsa1012065.

Zarin, D. A., T. Tse, R. J. Williams, and T. Rajakannan, 2017. Update on trial registration 11 years after the ICMJE policy was established. *The New England Journal of Medicine* 376 (4):383–91. doi: 10.1056/NEJMsr1601330.