Theory and performance of substitution models for estimating relative causal effects in nutritional epidemiology

Georgia D Tomova,[1,2,3] Mark S Gilthorpe,[1,3,4] and Peter WG Tennant[1,2,3]

[1]Leeds Institute for Data Analytics, University of Leeds, Leeds, United Kingdom; [2]Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom; [3]The Alan Turing Institute, London, United Kingdom; and [4]Obesity Institute, Leeds Beckett University, Leeds, United Kingdom

ABSTRACT

Background: Estimating relative causal effects (i.e., "substitution effects") is a common aim of nutritional research. In observational data, this is usually attempted using 1 of 2 statistical modeling approaches: the leave-one-out model and the energy partition model. Despite their widespread use, there are concerns that neither approach is well understood in practice.

Objectives: We aimed to explore and illustrate the theory and performance of the leave-one-out and energy partition models for estimating substitution effects in nutritional epidemiology.

Methods: Monte Carlo data simulations were used to illustrate the theory and performance of both the leave-one-out model and energy partition model, by considering 3 broad types of causal effect estimands: *1)* direct substitutions of the exposure with a single component, *2)* inadvertent substitutions of the exposure with several components, and *3)* average relative causal effects of the exposure instead of all other dietary sources. Models containing macronutrients, foods measured in calories, and foods measured in grams were all examined.

Results: The leave-one-out and energy partition models both performed equally well when the target estimand involved substituting a single exposure with a single component, provided all variables were measured in the same units. Bias occurred when the substitution involved >1 substituting component. Leave-one-out models that examined foods in mass while adjusting for total energy intake evaluated obscure estimands.

Conclusions: Regardless of the approach, substitution models need to be constructed from clearly defined causal effect estimands. Estimands involving a single exposure and a single substituting component are typically estimated more accurately than estimands involving more complex substitutions. The practice of examining foods measured in grams or portions while adjusting for total energy intake is likely to deliver obscure relative effect estimands with unclear interpretations. *Am J Clin Nutr* 2022;00:1–10.

Keywords: nutritional epidemiology, substitution models, substitution analysis, estimand, causal inference, compositional data

Introduction

Dietary guidelines often recommend substituting certain nutrients or foods with healthier alternatives. For instance, the United Kingdom's Scientific Advisory Committee on Nutrition recommends substituting saturated fats with unsaturated fats, to help ensure that saturated fats do not contribute >10% to total energy (1). Other examples include recommendations to replace refined grains with whole grains (2), and sugars with complex carbohydrates (3). Substitutions like these are typically informed by the cumulative evidence from both randomized control trials and prospective cohort studies.

In experimental studies, food substitutions are examined by conducting isocaloric dietary interventions (4). In contrast, observational studies must rely on estimating substitution effects using statistical methods and 2 approaches have become common. The first, known as the leave-one-out model, is a variation of the "standard model" for energy intake adjustment (5). Conceived as a means to emulate isocaloric interventions, this approach involves adjusting for total energy intake and all dietary sources except those being substituted with the exposure. The second approach, known as the energy partition model, involves including all sources of total energy intake as model covariates and subsequently subtracting the regression coefficients to estimate the substitution (6, 7).

Despite many previous attempts to explain the theory of substitution modeling in nutritional epidemiology (5, 8–11), and although both the leave-one-out and energy partition models are common in practice (5), there are concerns that they are not well understood (5, 9, 10). This study hence aims to

introduce, examine, and illustrate the theory and performance of the different substitution modeling approaches in nutritional epidemiology using a causal inference perspective and a series of illustrative simulations.

Target estimands and modeling approaches

Although the interpretation of substitution models has been discussed extensively elsewhere (5, 8–11), none have formally defined the causal effect estimands that these models target. When a substitution is sought in compositional data, i.e., data where ≥2 component "part" variables sum to a total "whole" variable (12), the corresponding target estimand represents a type of relative causal effect (13). Substitution models thus aim to estimate the joint effects of increasing the intake of an exposure while simultaneously decreasing the intake of ≥1 substituting components to maintain the same total (14, 15).

Depending on their specification, both the leave-one-out model and energy partition model can be used to estimate a range of relative effect estimands.

The leave-one-out model.

Because the leave-one-out model adjusts for total energy intake, and total energy intake is a collider for all sources of energy intake, it introduces a dependency (i.e., substituting effect) between all sources of energy intake that are not controlled in addition (13, 14). If adjustment is made only for the exposure and total energy intake, then the model estimates the average relative causal effect (i.e., the relative causal effect of the exposure instead of all other components) because all other components are free to contribute to keeping the total fixed (14, 15). For more specific substitutions, ≥1 components may be adjusted in addition to prevent them from contributing, thereby removing them from the substitution. Typically, a single component is left unadjusted (i.e., "left out" of the model) to estimate the relative causal effect of X_1 instead of X_2, where X_1 is the exposure and X_2 is the excluded component, giving the model its name (5). Substitutions between the exposure and >1 component may also be estimated by leaving ≥2 components unadjusted.

The leave-one-out model may be used to estimate isocaloric substitution effects for both macronutrients and foods, providing all components are measured in calories and sum to the total energy intake. Where foods are measured in grams or portions, they will not be compositional with total energy. Equal-mass substitution effects may instead be estimated by adjusting for the total food intake in grams or portions, respectively, but "mixed unit" models that adjust for total energy while examining foods in grams or portions will evaluate obscure estimands. Unfortunately, it is common for investigators to adjust for total energy intake without understanding the estimands that are produced (10, 14).

In practice, most leave-one-out models may be inadvertent, i.e., built without the specific intention of estimating a relative causal effect. Adjustment for total energy intake is practiced for a variety of reasons (8), without full appreciation of the impact on the estimand. For example, it is common for researchers to adjust for total energy alongside a selection of macronutrients or foods deemed "important confounders" without reference to a target substitution estimand (16–20).

The energy partition model and all-components model.

Although mathematically similar to the leave-one-out model, the energy partition model is philosophically distinct because it does not directly output substitution effects (7, 21). Instead, total energy intake is partitioned into constituent components, which are included as model covariates. The relative causal effect of X_1 instead of X_2 is then estimated from the difference between the β-coefficients for the exposure and the substituting component. Provided both components are measured in calories, this will be an isocaloric substitution. Alternatively, the average relative causal effect may be estimated by including and comparing terms for the exposure and a summary term for all "remaining energy intake" (i.e., the total energy intake from all components except the exposure). Bespoke substitutions can also be obtained by retaining all components, and explicitly calculating and comparing the weighted effects. For example, the average relative causal effect could be calculated from a model with all components by comparing the coefficient of the exposure with a weighted sum of the coefficients for all other components. Although technically a form of the energy partition model, we have previously named this latter approach the all-components model to differentiate it from other forms of the energy partition model (14). Like the leave-one-out model, the energy partition model can also be used with foods in grams to estimate equal-mass substitutions, provided the exposure and substituting components are all measured in grams.
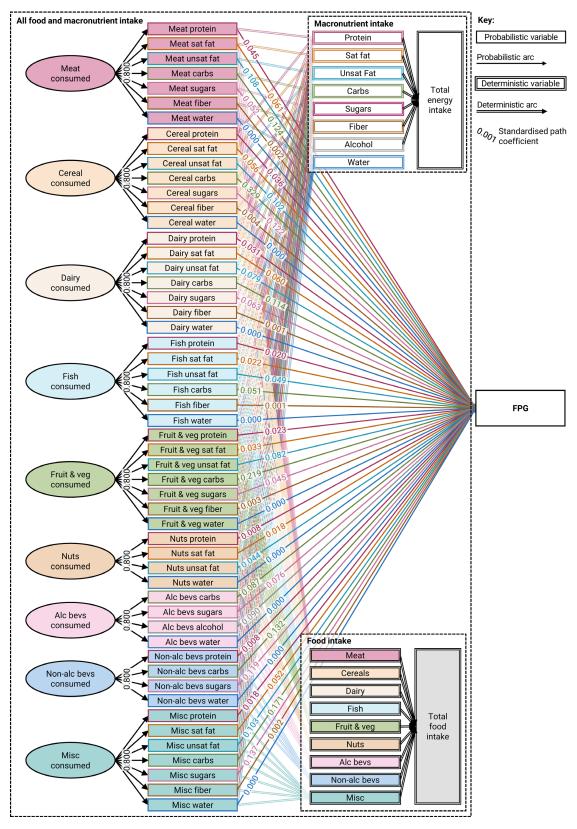
Methods

Illustrative example

To illustrate the theory and performance of each approach in estimating different relative (substitution) causal effects, we considered the effects of a macronutrient exposure, nonmilk extrinsic sugars (or "sugars"; kcal), and a food exposure, meat (kcal and g), on fasting plasma glucose concentration (FPG). The choice of exposures and outcome was arbitrary, and the presented causal effects were chosen for illustration.

Data simulation

Standardized data were simulated with the R package "dagitty" version 0.3-1 (22) based on a prespecified data generation process depicted in **Figure 1**. The simulated data reflect a scenario in which total energy intake is fully determined by the energy intake from 7 macronutrients and total food intake is fully determined by the intake from 9 food groups. The 7 macronutrients were *1)* sugars, *2)* complex carbohydrates, *3)* fiber, *4)* saturated fat, *5)* unsaturated fat, *6)* protein, and *7)* alcohol; and the 9 food groups were *1)* cereals, *2)* dairy, *3)* meat, *4)* fish, *5)* fruits and vegetables, *6)* nuts, *7)* alcoholic beverages, *8)* nonalcoholic beverages, and *9)* miscellaneous. Each macronutrient was assigned a unique effect on FPG using different path coefficients. Target means and SDs for each macronutrient (kcal) were informed by the United Kingdom's National Diet and Nutrition Survey (23), with SDs capped to ensure convergence (see **Supplemental Table 1**). Component macronutrient variables, and component water variables, were then simulated for each food and their means and SDs were scaled to match the distribution observed in the

FIGURE 1 Parametric directed acyclic graph of the simulated data generating process, showing the standardized path coefficients used in the simulation. This directed acyclic graph uses the notation introduced in Arnold et al. (13), which depicts deterministic variables and deterministic relations using double-outlined rectangles and double-lined arrows, respectively. Deterministic variables are variables that are fully determined by their parent variables, e.g., total energy intake is fully determined by the energy intake from the 7 macronutrients, which are in turn fully determined by the energy intake from the 45 component variables. All deterministic variables were derived from their parent variables; the path coefficients for deterministic relations are therefore not shown. FPG, fasting plasma glucose.

National Diet and Nutrition Survey (**Supplemental Table 2**) and the target means and SDs for the total intake of each macronutrient (Supplemental Table 1). Food intakes in calories were derived by summing the constituent macronutrients, and food intakes in grams were derived by adding the constituent macronutrients multiplied by their caloric density, with water added to match the moisture content observed in the National Diet and Nutrition Survey. Total energy intake (i.e., total calories consumed from all dietary sources), total food intake (i.e., total mass of food consumed from all dietary sources), remaining energy intake (i.e., total calories consumed from all dietary sources except the exposure), and remaining food intake (i.e., total food consumed from all dietary sources except the exposure) were calculated directly from the simulated variables to reflect the deterministic data generation process. In total, 100,000 simulated data sets were created, each comprising 1000 observations. We report the median effect estimate and the 2.5th and 97.5th centiles [representing 95% simulation interval (SI)] for each investigated model. To aid illustration, effect estimates are expressed in mg/dL per 100 kcal of nutrients or foods consumed (i.e., mg $\cdot$ dL^{-1} $\cdot$ 100 kcal^{-1}) and mg/dL per 100 g of foods consumed (i.e., mg $\cdot$ dL^{-1} $\cdot$ 100 g^{-1}).

Estimands and models examined

We examined 3 types of causal effect estimands across the 3 dietary unit scenarios of macronutrients (kcal), foods (kcal), and foods (g). The first group of estimands involved a direct substitution of the exposure for a single component. For macronutrients this was the relative causal effect of sugars instead of protein (R_1), and for foods this was the relative causal effect of meat instead of fish (kcal, R_4; and g, R_7). The second group of estimands involved a more complex substitution of the exposure for several components, as might arise from inadvertently adjusting for total energy intake without specifically intending to estimate a relative causal effect. For macronutrients, we considered the (inadvertent) estimand created by a model that adjusted for carbohydrates, alcohol, and total energy viewed as "important confounders"; this was the (inadvertent) relative causal effect of sugars instead of protein, total fat, and fiber (R_2). For foods, we considered the (inadvertent) estimands created by models that adjusted for fruit and veg, alcoholic beverages, nonalcoholic beverages, and total energy or total food; these were the (inadvertent) relative causal effects of meat instead of cereal, dairy, fish, nuts, and miscellaneous food in calories (R_5) and grams (R_8). The third group of estimands involved a substitution of the exposure for all other dietary components, i.e., average relative causal effects. For macronutrients this was the average relative causal effect of sugars (R_3) and for foods this was the average relative causal effect of meat (kcal, R_6; and g, R_9). Each estimand was evaluated using a "same unit" leave-one-out model, a comprehensive energy partition (i.e., all-components) model, and, where relevant, a simplified energy partition model and a "mixed unit" leave-one-out model. "Same unit" leave-one-out models are isocaloric and equal-mass models where the components are measured in calories and grams, respectively, and adjustment is made for total energy intake and total food intake, respectively. "Mixed unit" leave-one-out models are obscure models where the components are measured in grams and adjustment is made for total energy intake. We compared the estimates obtained from each model

either with the simulated truth (for the macronutrient models) or with estimates derived from the weighted average of the macronutrients making up each food (for the food group models). The **Supplemental Methods** provide further details of the estimands and models.

Results

Table 1 summarizes the results of the 24 models.

Macronutrient models

Both the "same unit" leave-one-out and (comprehensive) energy partition (i.e., all-components) models returned the true causal effect when the substitution involved a single macronutrient ($\widehat{R_{1.1}} = \widehat{R_{1.2}} = R_1 = 3.20$ mg/dL/100 kcal, i.e., the causal effect of 100 kcal from sugar on FPG was estimated to be 3.20 mg/dL larger than the effect of 100 kcal coming from protein, for the same total energy intake).

Where the substitution involved >1 macronutrient, the all-components model again returned the expected true value ($\widehat{R_{2.2}} = R_2 = 2.50$ mg/dL/100 kcal (i.e., the causal effect of 100 kcal from sugar on FPG was estimated to be 2.50 mg/dL larger than the average effect of 100 kcal coming from protein, fat, and fiber, for the same total energy intake), but the "same unit" leave-one-out model returned a slightly biased smaller value ($\widehat{R_{2.1}} = 2.39$ mg/dL/100 kcal).

The all-components model returned the expected true value for the average relative causal effect of sugar ($\widehat{R_{3.3}} = R_3 = 1.75$ mg/dL/100 kcal, i.e., the causal effect of 100 kcal from sugar on FPG was estimated to be 1.75 mg/dL larger than the average effect of 100 kcal coming from all other macronutrients, for the same total energy intake), but both the "same unit" leave-one-out model and the simple energy partition model returned the same slightly biased larger value ($\widehat{R_{3.1}} = \widehat{R_{3.2}} = 2.05$ mg/dL/100 kcal).

Food group models—in calories

Both the "same unit" leave-one-out and (comprehensive) energy partition (i.e., all-components) models returned the same value when the substitution involved a single food ($\widehat{R_{4.1}} = \widehat{R_{4.2}} = 0.20$ mg/dL/100 kcal, i.e., the causal effect of 100 kcal from meat on FPG was estimated to be 0.20 mg/dL larger than the effect of 100 kcal coming from fish, for the same total energy intake), which was similar to but slightly larger than the expected true value ($R_4 = 0.16$ mg/dL/100 kcal).

Where the substitution involved >1 variable, both the "same unit" leave-one-out and all-components models returned similar values ($\widehat{R_{6.1}} = -0.37$ mg/dL/100 kcal and $\widehat{R_{5.2}} = -0.35$ mg/dL/100 kcal, respectively, i.e., the causal effect of 100 kcal from meat on FPG was estimated to be 0.37 or 0.35 mg/dL smaller than the average effect of 100 kcal coming from cereal, dairy, fish, nuts, and miscellaneous foods, for the same total energy intake), both of which were smaller than and almost half the expected true value ($R_5 = -0.65$ mg/dL/100 kcal).

The all-components model returned an average relative causal effect of meat ($\widehat{R_{6.3}} = -0.47$ mg/dL/100 kcal, i.e., the causal effect of 100 kcal from meat on FPG was estimated to be 0.47 mg/dL smaller than the average effect of 100 kcal coming from all other foods, for the same total energy intake) that was smaller than and almost half the expected true value (R_6

TABLE 1 Full details of the 9 estimands and 24 models examined in the simulated data[1]

Estimand symbol	Estimand Name	Simulated or derived[2] value	Estimate symbol	Model Name	Model Formula	Estimated value (95% SI)
R_1	Relative causal effect of sugars instead of protein	3.20^3	$\widehat{R_{1.1}}$	"Same unit" leave-one-out	$\widehat{\mathrm{FPG}} = \widehat{a_0} + \widehat{a_1}\text{sugars} + \widehat{a_2}\text{carbs} + \widehat{a_3}\text{unsat fat} + \widehat{a_4}\text{sat fat} + \widehat{a_5}\text{fiber} + \widehat{a_6}\text{alcohol} + \widehat{a_7}\text{TE} + \varepsilon$ $\widehat{R_{1.1}} = \widehat{a_1}$	3.20^3 (2.52, 3.88)
			$\widehat{R_{1.2}}$	(Comprehensive) energy partition/all-components	$\widehat{\mathrm{FPG}} = \widehat{b_0} + \widehat{b_1}\text{sugars} + \widehat{b_2}\text{protein} + \widehat{b_3}\text{carbs} + \widehat{b_4}\text{unsat fat} + \widehat{b_5}\text{sat fat} + \widehat{b_6}\text{fiber} + \widehat{b_7}\text{alcohol} + \varepsilon$ $\widehat{R_{1.2}} = \widehat{b_1} - \widehat{b_2}$	3.20^3 (2.52, 3.88)
R_2	"Inadvertent" relative causal effect of sugars instead of protein, total fat, and fiber	2.50^3	$\widehat{R_{2.1}}$	"Same unit" leave-one-out	$\widehat{\mathrm{FPG}} = \widehat{c_0} + \widehat{c_1}\text{sugars} + \widehat{c_2}\text{TE} + \widehat{c_3}\text{carbs} + \widehat{c_4}\text{alcohol} + \varepsilon$ $\widehat{R_{2.1}} = \widehat{c_1}$	2.39^3 (1.92, 2.85)
			$\widehat{R_{2.2}}$	All-components	$\widehat{\mathrm{FPG}} = \widehat{b_0} + \widehat{b_1}\text{sugars} + \widehat{b_2}\text{protein} + \widehat{b_3}\text{carbs} + \widehat{b_4}\text{unsat fat} + \widehat{b_5}\text{sat fat} + \widehat{b_6}\text{fiber} + \widehat{b_7}\text{alcohol} + \varepsilon$ $\widehat{R_{2.2}} = \widehat{b_1} - \widehat{b_{2,4,5,6}}$ where $\widehat{b_{2,4,5,6}} = \sum iw_ib_i$ and $i = \{2, 4, 5, 6\}$	2.50^3 (2.02, 2.98)
R_3	Average relative causal effect of sugar	1.75^3	$\widehat{R_{3.1}}$	"Same unit" leave-one-out	$\widehat{\mathrm{FPG}} = \widehat{d_0} + \widehat{d_1}\text{sugars} + \widehat{d_2}\text{TE} + \varepsilon$ $\widehat{R_{3.1}} = \widehat{d_1}$	2.05^3 (1.57, 2.52)
			$\widehat{R_{3.2}}$	(Simple) energy partition	$\widehat{\mathrm{FPG}} = \widehat{f_0} + \widehat{f_1}\text{sugars} + \widehat{f_2}\text{RE} + \varepsilon$ $\widehat{R_{3.2}} = \widehat{f_1} - \widehat{f_2}$	2.05^3 (1.57, 2.52)
			$\widehat{R_{3.3}}$	(Comprehensive) energy partition/all-components	$\widehat{\mathrm{FPG}} = \widehat{b_0} + \widehat{b_1}\text{sugars} + \widehat{b_2}\text{protein} + \widehat{b_3}\text{carbs} + \widehat{b_4}\text{unsat fat} + \widehat{b_5}\text{sat fat} + \widehat{b_6}\text{fiber} + \widehat{b_7}\text{alcohol} + \varepsilon$ $\widehat{R_{3.3}} = \widehat{b_1} - \widehat{b_{2:7}}$ where $\widehat{b_{2:7}} = \sum iw_ib_i$ and $i = \{2, 3, 4, 5, 6, 7\}$	1.75^3 (1.29, 2.21)
R_4	Relative causal effect of meat instead of fish (kcal)	0.16^3	$\widehat{R_{4.1}}$	"Same unit" leave-one-out	$\widehat{\mathrm{FPG}} = \widehat{g_0} + \widehat{g_1}\text{meat} + \widehat{g_2}\text{cereal} + \widehat{g_3}\text{dairy} + \widehat{g_4}\text{fruit veg} + \widehat{g_5}\text{nuts} + \widehat{g_6}\text{alc bev} + \widehat{g_7}\text{nonalc bev} + \widehat{g_8}\text{misc} + \widehat{g_9}\text{TE} + \varepsilon$ $\widehat{R_{4.1}} = \widehat{g_1}$	0.20^3 (−0.21, 0.61)
			$\widehat{R_{4.2}}$	(Comprehensive) energy partition/all-components	$\widehat{\mathrm{FPG}} = \widehat{h_0} + \widehat{h_1}\text{meat} + \widehat{h_2}\text{cereal} + \widehat{h_3}\text{dairy} + \widehat{h_4}\text{fish} + \widehat{h_5}\text{fruit veg} + \widehat{h_6}\text{nuts} + \widehat{h_7}\text{alc bev} + \widehat{h_8}\text{nonalc bev} + \widehat{h_9}\text{misc} + \varepsilon$ $\widehat{R_{4.2}} = \widehat{h_1} - \widehat{h_4}$	0.20^3 (−0.21, 0.61)
R_5	"Inadvertent" relative causal effect of meat instead of cereal, dairy, fish, nuts, and misc. (kcal)	-0.65^3	$\widehat{R_{5.1}}$	(Inadvertent) "same unit" leave-one-out	$\widehat{\mathrm{FPG}} = \widehat{i_0} + \widehat{i_1}\text{meat} + \widehat{i_2}\text{fruit veg} + \widehat{i_3}\text{alc bev} + \widehat{i_4}\text{nonalc bev} + \widehat{i_5}\text{TE} + \varepsilon$ $\widehat{R_{5.1}} = \widehat{i_1}$	-0.37^3 (−0.52, −0.20)
			$\widehat{R_{5.2}}$	All-components	$\widehat{\mathrm{FPG}} = \widehat{h_0} + \widehat{h_1}\text{meat} + \widehat{h_2}\text{cereal} + \widehat{h_3}\text{dairy} + \widehat{h_4}\text{fish} + \widehat{h_5}\text{fruit veg} + \widehat{h_6}\text{nuts} + \widehat{h_7}\text{alc bev} + \widehat{h_8}\text{nonalc bev} + \widehat{h_9}\text{misc} + \varepsilon$ $\widehat{R_{5.2}} = \widehat{h_1} - \widehat{h_{2,3,4,6,9}}$ where $h_{2,3,4,6,9} = \sum iw_ih_i$ and $i = \{2, 3, 4, 6, 9\}$	-0.35^3 (−0.52, −0.19)

(Continued)

TABLE 1 (*Continued*)

Estimand symbol	Name	Simulated or derived[2] value	Estimate symbol	Name	Model — Formula	Estimated value (95% SI)
R_6	Average relative causal effect of meat (kcal)	-0.78^3	$\widehat{R_{6.1}}$	"Same unit" leave-one-out	$\widehat{FPG} = \widehat{j_0} + \widehat{j_1}\text{meat} + \widehat{j_2}\text{TE} + \varepsilon$	-0.42^3 (−0.59, −0.25)
			$\widehat{R_{6.2}}$	(Simple) energy partition	$\widehat{R_{6.1}} = \widehat{j_1}$ $\widehat{FPG} = \widehat{k_0} + \widehat{k_1}\text{meat} + \widehat{k_2}\text{RE} + \varepsilon$	-0.42^3 (−0.59, −0.25)
			$\widehat{R_{6.3}}$	(Comprehensive) energy partition/all-components	$\widehat{R_{6.2}} = \widehat{k_1} - \widehat{k_2}$ $\widehat{FPG} = \widehat{h_0} + \widehat{h_1}\text{meat} + \widehat{h_2}\text{cereal} + \widehat{h_3}\text{dairy} + \widehat{h_4}\text{fish} + \widehat{h_5}\text{fruit veg} + \widehat{h_6}\text{nuts} + \widehat{h_7}\text{alc bev} + \widehat{h_8}\text{nonalc bev} + \widehat{h_9}\text{misc} + \varepsilon$ $\widehat{R_{6.3}} = \widehat{h_1} - \widehat{h_{2:9}}$ where $\widehat{h_{2:9}} = \sum iw_i h_i$ and $i = \{2, 3, 4, 5, 6, 7, 8, 9\}$	-0.47^3 (−0.63, −0.30)
R_7	Relative causal effect of meat instead of fish (g)	5.15^4	$\widehat{R_{7.1}}$	"Same unit" leave-one-out (with total food intake)	$\widehat{FPG} = \widehat{l_0} + \widehat{l_1}\text{meat} + \widehat{l_2}\text{cereal} + \widehat{l_3}\text{dairy} + \widehat{l_4}\text{fruit veg} + \widehat{l_5}\text{nuts} + \widehat{l_6}\text{alc bev} + \widehat{l_7}\text{nonalc bev} + \widehat{l_8}\text{misc} + \widehat{l_9}\text{TF} + \varepsilon$ $\widehat{R_{7.1}} = \widehat{l_1}$	4.81^4 (3.23, 6.32)
			$\widehat{R_{7.2}}$	"Mixed unit" leave-one-out (with total energy)	$\widehat{FPG} = \widehat{m_0} + \widehat{m_1}\text{meat} + \widehat{m_2}\text{cereal} + \widehat{m_3}\text{dairy} + \widehat{m_4}\text{fruit veg} + \widehat{m_5}\text{nuts} + \widehat{m_6}\text{alc bev} + \widehat{m_7}\text{nonalc bev} + \widehat{m_8}\text{misc} + \widehat{m_9}\text{TE} + \varepsilon$ $\widehat{R_{7.2}} = \widehat{m_1}$	-0.94^4 (−1.58, −0.31)
			$\widehat{R_{7.3}}$	All-components	$\widehat{FPG} = \widehat{n_0} + \widehat{n_1}\text{meat} + \widehat{n_2}\text{cereal} + \widehat{n_3}\text{dairy} + \widehat{n_4}\text{fish} + \widehat{n_5}\text{fruit veg} + \widehat{n_6}\text{nuts} + \widehat{n_7}\text{alc bev} + \widehat{n_8}\text{nonalc bev} + \widehat{n_9}\text{misc} + \varepsilon$ $\widehat{R_{7.3}} = \widehat{n_1} - \widehat{n_4}$	4.81^4 (3.23, 6.32)
R_8	"Inadvertent" relative causal effect of meat instead of cereal, dairy, fish, nuts, and misc. (g)	1.44^4	$\widehat{R_{8.1}}$	(Inadvertent) "same unit" leave-one-out (with total food intake)	$\widehat{FPG} = \widehat{o_0} + \widehat{o_1}\text{meat} + \widehat{o_2}\text{fruit veg} + \widehat{o_3}\text{alc bev} + \widehat{o_4}\text{nonalc bev} + \widehat{o_5}\text{TF} + \varepsilon$ $\widehat{R_{8.1}} = \widehat{o_1}$	4.60^4 (3.21, 6.06)
			$\widehat{R_{8.2}}$	(Inadvertent) "mixed unit" leave-one-out (with total energy)	$\widehat{FPG} = \widehat{p_0} + \widehat{p_1}\text{meat} + \widehat{p_2}\text{fruit veg} + \widehat{p_3}\text{alc bev} + \widehat{p_4}\text{nonalc bev} + \widehat{p_5}\text{TE} + \varepsilon$ $\widehat{R_{8.2}} = \widehat{p_1}$	-1.11^4 (−1.64, −0.60)
			$\widehat{R_{8.3}}$	All-components	$\widehat{FPG} = \widehat{n_0} + \widehat{n_1}\text{meat} + \widehat{n_2}\text{cereal} + \widehat{n_3}\text{dairy} + \widehat{n_4}\text{fish} + \widehat{n_5}\text{fruit veg} + \widehat{n_6}\text{nuts} + \widehat{n_7}\text{alc bev} + \widehat{n_8}\text{nonalc bev} + \widehat{n_9}\text{misc} + \varepsilon$ $\widehat{R_{8.3}} = \widehat{n_1} - \widehat{n_{2,3,4,6,9}}$ where $n_{2,3,4,6,9} = \sum iw_i n_i$ and $i = \{2, 3, 4, 6, 9\}$	5.42^4 (4.51, 6.33)
R_9	Average relative causal effect of meat (g)	6.48^4	$\widehat{R_{9.1}}$	"Same unit" leave-one-out (with total food intake)	$\widehat{FPG} = \widehat{q_0} + \widehat{q_1}\text{meat} + \widehat{q_2}\text{TF} + \varepsilon$ $\widehat{R_{9.1}} = \widehat{q_1}$	8.14^4 (6.55, 9.80)

(*Continued*)

TABLE 1 (*Continued*)

Estimand			Model			Estimated value (95% SI)
Name	Estimand symbol	Simulated or derived[2] value	Estimate symbol	Name	Formula	
			$\widehat{R_{9.2}}$	"Mixed unit" leave-one-out (with total energy)	$\widehat{FPG} = \hat{r}_0 + \hat{r}_1\text{meat} + \hat{r}_2\text{TE} + \varepsilon$	-1.17^4 $(-1.70, -0.66)$
			$\widehat{R_{9.3}}$	(Simple) energy partition	$\widehat{R_{9.2}} = \hat{r}_1 - \hat{r}_2$ $\widehat{FPG} = \hat{s}_0 + \hat{s}_2\text{meat} + \hat{s}_2\text{RF} + \varepsilon$	8.14^4 $(6.55, 9.80)$
			$\widehat{R_{9.4}}$	All-components	$\widehat{R_{9.3}} = \hat{s}_1 - \hat{s}_2$ $\widehat{FPG} = \hat{n}_0 + \hat{n}_1\text{meat} + \hat{n}_2\text{cereal} + \hat{n}_3\text{dairy} + \hat{n}_4\text{fish} + \hat{n}_5\text{fruit veg} + \hat{n}_6\text{nuts} + \hat{n}_7\text{alc bev} + \hat{n}_8\text{nonalc bev} + \hat{n}_9\text{misc} + \varepsilon$ $\widehat{R_{9.4}} = \hat{n}_1 - \bar{n}_{2:9}$ where $\bar{n}_{2:9} = \sum iw_i \hat{n}_i$ and $i = \{2, 3, 4, 5, 6, 7, 8, 9\}$	6.06^4 $(5.12, 7.05)$

[1] FPG, fasting plasma glucose; RE, remaining energy intake; RF, remaining food (mass) intake; SI, simulation interval; TE, total energy intake; TF, total food (mass) intake.
[2] Estimated by directly calculating the weighted-average effect; for further details see the analytic code.
[3] Units are mg · dL^{-1} · 100 kcal^{-1}.
[4] Units are mg · dL^{-1} · 100 g^{-1}.

$= -0.78$ mg/dL/100 kcal). Both the "same unit" leave-one-out model and the simple energy partition model returned identical estimates that were slightly more biased ($\widehat{R_{6.1}} = \widehat{R_{6.2}} = -0.42$ mg/dL/100 kcal).

Food group models—in grams

Both the "same unit" leave-one-out and (comprehensive) energy partition (i.e., all-components) models returned the same value when the substitution involved a single food and adjustment was made for total food intake ($\widehat{R_{7.1}} = \widehat{R_{7.3}} = 4.81$ mg/dL/100 g, i.e., the causal effect of 100 g from meat on FPG was estimated to be 4.81 mg/dL larger than the effect of 100 g coming from fish, for the same total food intake), which was slightly smaller than the expected true value ($R_7 = 5.15$ mg/dL/100 g). The "mixed unit" leave-one-out-model that adjusted for total energy intake returned a radically different value with the opposite sign to both the isocaloric and equal-mass estimands ($\widehat{R_{7.2}} = -0.94$ mg/dL/100 g, i.e., the causal effect of 100 g from meat on FPG was estimated to be 0.94 mg/dL smaller than the effect of 100 g coming from fish, for the same total energy intake).

Where the substitution involved >1 variable, both the "same unit" leave-one-out and all-components models returned similar values ($\widehat{R_{8.1}} = 4.60$ mg/dL/100 g and $\widehat{R_{8.3}} = 5.42$ mg/dL/100 g, respectively, i.e., the causal effect of 100 g from meat on FPG was estimated to be 4.60 or 5.42 mg/dL larger than the average effect of 100 g coming from cereal, dairy, fish, nuts, and miscellaneous foods, for the same total food intake), which were much larger than the derived estimate ($R_8 = 1.44$ mg/dL/100 g). The "mixed unit" leave-one-out-model that adjusted for total energy intake returned a radically different value of equal sign to the isocaloric estimand and opposite sign to the equal-mass estimand ($\widehat{R_{8.2}} = -1.11$ mg/dL/100 g, i.e., the causal effect of 100 g from meat on FPG was estimated to be 1.11 mg/dL smaller than the average effect of 100 g coming from cereal, dairy, fish, nuts, and miscellaneous foods, for the same total energy intake).

The all-components model returned a slightly smaller average relative causal effect of meat than the expected true value ($\widehat{R_{9.4}} = 6.06$ mg/dL/100 g compared with $R_9 = 6.48$ mg/dL/100 g, i.e., the true causal effect of 100 g from meat on FPG was 6.48 mg/dL larger than the average effect of 100 g coming from all other foods, for the same total food intake). Both the "same unit" leave-one-out model and the simple energy partition model returned identical estimates that were larger than the expected true values ($\widehat{R_{9.1}} = \widehat{R_{9.3}} = 8.14$ mg/dL/100 g). The "mixed unit" leave-one-out-model that adjusted for total energy intake returned a radically different value of equal sign to the isocaloric estimand and opposite sign to the equal-mass estimand ($\widehat{R_{9.2}} = -1.17$ mg/dL/100 g, i.e., the causal effect of 100 g from meat on FPG was estimated to be 1.17 mg/dL smaller than the average effect of 100 g coming from all other foods, for the same total energy intake).

Discussion

Principal findings

This study examined and compared the performance of the 2 most common approaches to estimating relative (i.e., substitution) effects in nutritional epidemiology. We demonstrated

that "same unit" leave-one-out models and energy partition models perform equally well when the target estimand involves substituting a single exposure with a single component. Once the substitution involves ≥2 components, both models can produce biased estimates that diverge from the expected true weighted average effects.

We have also shown that isocaloric and equal-mass substitution estimates can be reliably estimated using either model provided all variables are measured in the same units. "Mixed unit" leave-one-out models that examine foods in mass while adjusting for total energy intake evaluate obscure estimands that may be sign discordant from both the isocaloric and equal-mass estimands.

Analysis and explanation

In substitution modeling, the estimand is determined jointly by the exposure and by the dietary components chosen for substitution. There is hence no single "relative causal effect" but an array of potential estimands that may be considered, each requiring appropriate modeling and interpretation. Because nutritional data are compositional, the causal effect being estimated by a particular model may depend not only on the components in the model, but also on those that are absent, whether purposefully or inadvertently. This is particularly true where the "total" is conditioned on, because a relative effect will be established with all components that are absent from the model (13).

Our simulations show the hazards of failing to consider the estimand being sought, particularly when adjusting for total energy intake. In both the isocaloric and equal-mass models of meat compared with fish, a positive effect was observed. However, when the comparison was with cereal, dairy, fish, nuts, and miscellaneous foods, the effect differed depending on whether the substitution concerned calories or grams. Unfortunately, simply adjusting for total energy intake does not guarantee the substitution is isocaloric; the dietary components must also be measured and analyzed in calories to ensure the data are compositional. Where different units are used, the relation between the components and the total is no longer compositional, resulting in an obscure estimand that may be extremely misleading. Our "mixed unit" leave-one-out models of meat and other food components in grams that adjusted for total energy intake returned ambiguous coefficients with opposite sign to the appropriate equal-mass estimands and, on 1 occasion, the appropriate isocaloric estimand. We were unable to establish a firm pattern or meaning for these estimands and therefore advise that the coefficients from "mixed unit" leave-one-out models be treated with caution.

Even where specified appropriately, we have shown that both the "same unit" leave-one-out and energy partition models are prone to bias when the substitution involves ≥2 components. We have previously shown that the standard model for energy intake adjustment, which adjusts for total energy, does not robustly estimate the average relative causal effect owing to "composite variable bias" (14). This occurs because different components with different effects are combined into a single "total energy intake" variable that captures an imperfect average of the individual components. As an extension of the standard model, the leave-one-out model suffers the same problem with

substitutions involving >1 component, but the energy partition model also suffers the same issue where composite terms, such as "remaining energy intake," are used. Where the model includes all components, this problem is mitigated because the weighted average effect can be calculated from each part. However, even this approach is not guaranteed to produce a perfect weighted average because all variables, to a greater or lesser extent, are composites of smaller parts. This explains why none of the food models returned the same estimate as the derived weighted average effect; every food contained multiple macronutrient components with different effects.

Implications and recommendations

Substitution effects include ≥2 variables in the exposure regime, making it particularly necessary to ensure the estimand is well-defined. We recommend that all studies seeking substitution effects should clearly state their target estimand and justify the adjustment strategy accordingly; directed acyclic graphs can be particularly useful in this regard (24).

What constitutes a well-defined intervention in experimental research and public health practice may differ from what constitutes a well-defined estimand in observational data. In experimental and interventional contexts, greater focus is typically given to food groups or portions because these represent the units of consumption. Food groups or portions may equally be examined in observational data, but their composite nature means that the estimated effects may be less consistent than for macronutrient variables. We encourage researchers to consider which units are most appropriate for their specific interests and recognize the trade-offs involved.

Wherever ≥2 components are involved in the substitution, there is scope for composite variable bias unless the individual effects are estimated and combined using an all-components approach. Because the leave-one-out approach is more common in the literature than the energy partition model (16–20), this raises concerns about the validity of existing estimates that involve more complex substitutions.

Imperfect substitution models may nevertheless represent less of a concern than inadvertent substitution models that arise from naïve adjustments for total energy intake. Total energy intake is routinely included in nutrition epidemiology models for a variety of reasons (8). Unfortunately, if the other model variables are not carefully selected with a specific estimand in mind, the model may estimate inadvertent or obscure causal effects. Where isocaloric substitution effects are desired, all dietary variables must be measured and analyzed in calorie units, regardless of whether the leave-one-out or energy partition model is used and regardless of whether the variables are macronutrients, foods in grams, or foods in portions. Equal-mass substitutions may also be estimated using either approach by including mass variables whether in grams or portions; for the leave-one-out model this will require adjusting for total food in grams or portions, respectively.

These findings have clear implications, not just for the conduct of future research, but for the appraisal and synthesis of existing studies. Meta-analyses of dietary exposures rarely separate studies based on their modeling strategy, making their results difficult to interpret. Where estimates from different estimands have been combined, the results may be misleading

or meaningless. Relative causal effects are particularly difficult to compare or summarize because it is extremely common to adjust for several components in addition to total energy intake, creating unique substitution effects (16–20). Studies that have considered ≥1 food variables in grams while adjusting for total energy may contribute obscure estimates with opposite signs to those estimated from appropriate isocaloric or equal-mass models. We believe this is likely to explain some of the apparent heterogeneity in effect estimates between studies, and we recommend that meta-analyses should only pool estimates of the same estimand.

Limitations and caveats

Our study used simulated data that oversimplify reality. Although informed by real data, our primary aim was to demonstrate the methodological topics at hand; the effects therefore should not be interpreted as true effects. All variables were simulated as multivariate normal, with no measurement error or interactions, which does not reflect reality. All variables were transformed toward observed means (23), but SDs had to be reduced and some negative and/or biologically implausible values were simulated.

We did not simulate or examine the influence of measurement error. Although important, the issues with measurement error are distinct from those presented in the current article. Similarly, we did not simulate or examine the influence of confounding. Although confounding will be ubiquitous in practice, there is no specific relevance to substitution modeling beyond what we have shown previously (14). Both the leave-one-out and energy partition models will be equally robust to confounding provided the substitution involves only 1 component; but residual confounding may arise where ≥2 components are unadjusted (14).

Conclusion

The "leave-one-out" and energy partition models perform equally well when estimating relative effects that involve substituting a single exposure with a single dietary component. Where more components are involved, both approaches may return biased estimates. Regardless of the approach, substitution models need to be constructed from clearly defined causal effect estimands. The practice of examining foods measured in grams or portions while adjusting for total energy intake is likely to deliver obscure relative effect estimands with unclear interpretations.

Data Availability

Analytic code, which includes data simulations, is publicly and freely available without restriction at https://github.com/georgia tomova/nutrition-substitution.

References

1. Scientific Advisory Committee on Nutrition (SACN). Saturated fats and health. [Internet]. London (United Kingdom): Public Health England; 2019. [Accessed 2021 Dec 1]. Available from: https://assets.p ublishing.service.gov.uk/government/uploads/system/uploads/attachm ent_data/file/814995/SACN_report_on_saturated_fat_and_health.pdf.
2. US Department of Agriculture and US Department of Health and Human Services (US DHHS). Dietary Guidelines for Americans, 2020–2025. [Internet]. 9th ed. Washington (DC): USDA and US DHHS; 2020. [Accessed 2021 Dec 1]. Available from: https://www.dietaryguide lines.gov.
3. Scientific Advisory Committee on Nutrition (SACN). Carbohydrates and health. [Internet]. London (United Kingdom): Public Health England; 2015. [Accessed 2021 Dec 1]. Available from: https://assets .publishing.service.gov.uk/government/uploads/system/uploads/attac hment_data/file/445503/SACN_Carbohydrates_and_Health.pdf.
4. Schwingshackl L, Neuenschwander M, Hoffmann G, Buyken AE, Schlesinger S. Dietary sugars and cardiometabolic risk factors: a network meta-analysis on isocaloric substitution interventions. Am J Clin Nutr 2020;111(1):187–96.
5. Song M, Giovannucci E. Substitution analysis in nutritional epidemiology: proceed with caution. Eur J Epidemiol 2018;33(2):137–40.
6. Willett W. Nutritional epidemiology. 3rd ed. New York: Oxford University Press; 2013.
7. Mekary RA, Willett WC, Hu FB, Ding EL. Isotemporal substitution paradigm for physical activity epidemiology and weight change. Am J Epidemiol 2009;170(4):519–27.
8. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. Am J Clin Nutr 1997;65(4):1220S–8S.
9. Ibsen DB, Jakobsen MU, Overvad K. Statistical models in nutritional epidemiology: more focus on the interpretation and argumentation for variable selection. Am J Clin Nutr 2019;110(6):1510.
10. Ibsen DB, Laursen ASD, Würtz AML, Dahm CC, Rimm EB, Parner ET, et al. Food substitution models for nutritional epidemiology. Am J Clin Nutr 2021;113(2):294–303.
11. Kipnis V, Freedman LS, Brown CC, Hartman A, Schatzkin A, Wacholder S. Interpretation of energy adjustment models for nutritional epidemiology. Am J Epidemiol 1993;137(12):1376–80.
12. Aitchison J. The statistical analysis of compositional data. J R Statist Soc B 1982;44(2):139–77.
13. Arnold KF, Berrie L, Tennant PWG, Gilthorpe MS. A causal inference perspective on the analysis of compositional data. Int J Epidemiol 2020;49(4):1307–13.
14. Tomova GD, Arnold KF, Gilthorpe MS, Tennant PWG. Adjustment for energy intake in nutritional research: a causal inference perspective. Am J Clin Nutr 2022;115(1):189–98.
15. Breskin A, Murray EJ. Commentary: compositional data call for complex interventions. Int J Epidemiol 2020;49(4):1314–5.
16. Chazelas E, Srour B, Desmetz E, Kesse-Guyot E, Julia C, Deschamps V, et al. Sugary drink consumption and risk of cancer: results from NutriNet-Santé prospective cohort. BMJ 2019;366(8207):l2408.
17. Zheng Y, Li Y, Satija A, Pan A, Sotos-Prieto M, Rimm E, et al. Association of changes in red meat consumption with total and cause specific mortality among US women and men: two prospective cohort studies. BMJ 2019;365(8203):l2110.
18. Srour B, Fezeu LK, Kesse-Guyot E, Allès B, Méjean C, Andrianasolo RM, et al. Ultra-processed food intake and risk of cardiovascular disease: prospective cohort study (NutriNet-Santé). BMJ 2019;365(8201):l1451.
19. Pacheco LS, Anderson CAM, Lacey JV Jr, Giovannucci EL, Lemus H, Araneta MRG, et al. Sugar-sweetened beverages and colorectal cancer risk in the California Teachers Study. PLoS One 2019;14(10):e0223638.
20. Katagiri R, Sawada N, Goto A, Yamaji T, Iwasaki M, Noda M, et al. Association of soy and fermented soy product intake with total and cause specific mortality: prospective cohort study. BMJ 2020;368(8231):m34.
21. Hu FB, Stampfer MJ, Rimm E, Ascherio A, Rosner BA, Spiegelman D, et al. Dietary fat and coronary heart disease: a comparison of approaches

for adjusting for total energy intake and modeling repeated dietary measurements. Am J Epidemiol 1999;149(6):531–40.

22. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package 'dagitty.' Int J Epidemiol 2016;45(6):1887–94.

23. NatCen Social Research, MRC Elsie Widdowson Laboratory. National Diet and Nutrition Survey Years 1–9, 2008/09–2016/17 [Internet]. 15th ed. Colchester (United Kingdom): UK Data Service; 2019. [Cited 2021 Aug 15]. Available from: https://doi.org/10.5255/UKDA-SN-6533-19.

24. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol 2021;50(2):620–32.