BAE
Bio-based and Applied Economics

AIEAA
ASSOCIAZIONE ITALIANA DI
ECONOMIA AGRARIA E APPLICATA

Citation: M. Mazzocchi, S. Capacci, B. Biondi (2022). Causal inference on the impact of nutrition policies using observational data. *Bio-based and Applied Economics* 11(1): 3-20. doi: 10.36253/bae-12411

Received: December 9, 2021
Accepted: May 29, 2022
Published: July 22, 2022

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Editor: Meri Raggi.
Discussant: Gianluca Stefani, Meri Raggi

ORCID
MM: 0000-0002-0498-8031
SC: 0000-0003-4162-682X
BB: 0000-0003-0343-1677

BAE 10th Anniversary papers

Causal inference on the impact of nutrition policies using observational data

Mario Mazzocchi*, Sara Capacci, Beatrice Biondi

Dipartimento di Scienze Statistiche, Università di Bologna
*Corresponding author. E-mail: m.mazzocchi@unibo.it

Abstract. We discuss the state-of-the-art in the application of quasi-experimental methods to estimate the impact of nutrition policies based on observational data. This field of application is less mature compared to other settings, especially labour and health policy, as food economists have started to implement widely counterfactual methods only over the last decade. We review the underlying assumptions behind the most prominent methods, when they can be regarded as credible and if/when they can be tested. We especially focus on the problem of dealing with unobserved confounding factors, emphasizing recent evidence on the limitations of propensity score methods, and the hard task of convincing reviewers about the quality of instrumental variables. We discuss the application of Difference-in-Difference, with an emphasis on its potential in consumer panel data applications, and how results from Regression Discontinuity Design studies should be interpreted. Finally, we cover the estimation of counterfactual outcomes using structural methods and provide an overview of recent developments and current gaps.

Keywords: quasi-experimental methods, policy evaluation, nutrition policy, assumptions.
JEL codes: C54, C21, Q18, I18.

1. INTRODUCTION

The call for evidence-based policy decision has generated an exponential growth in food policy evaluations over the last decade. Table 1 shows the counts obtained from a Google Scholar search for relevant keywords over the last three decades. Between 2011 and 2020 the number of hits for the generic term “Food Policy” is 4.8 times the baseline period, about 229,000 documents compared to 47,400 over the decade 1991-2000. Adding the keyword “impact evaluation” highlights a much faster trend. The increase in the number of Google Scholar hits is almost 25-fold. The proportion of papers with these keywords in relation to the simple “food policy” search results was only 0.63% in the 1990s, and rose to 3.24% in the 2010s. This pattern is confirmed by more specific keyword searches. For example, the additional keyword “causal identification” returns a 67-fold rise over two decades, and when looking for a specific method like “difference-in-difference”, hits grew from almost zero to 2,160, a 108-fold increase.

Firenze University Press | www.fupress.com/bae

Table 1. Food policy and evaluations in Google Scholar keywords searches over three decades.

	(a)		(b)		(c)		(c)/(a)	(c)/(b)
	1991-2000	%	2001-2010	%	2011-2020	%		
"Food policy"	47.4		178		229		4.83	1.29
"Food policy" and "impact evaluation"	0.3	0.63	1.89	1.06	7.43	3.24	24.77	3.93
"Food policy" and "randomized experiment"	0.02	0.04	0.46	0.26	1.86	0.81	93.00	4.04
"Food policy" and "counterfactual"	0.37	0.78	1.86	1.04	5.64	2.46	15.24	3.03
"Food policy" and "quasi-experimental methods"	0.006	0.01	0.07	0.04	0.36	0.16	60.00	5.14
"Food policy" and "causal identification"	0.003	0.01	0.009	0.01	0.2	0.09	66.67	22.22
"Food policy" and "difference-in-difference"	0.02	0.04	0.48	0.27	2.16	0.94	108.00	4.50

Source: Our search, Google Scholar accessed on 16/5/2022.

While this trend is similar in other areas of applied economics like health economics or energy economics, it has brought a small revolution in the agricultural economics field. In the year 2000, the journal Food Policy was 123th by impact factor within a population of 166 economics journal. In 2019 the journal ranked 28th out of 373 economics journals, and has been regularly the highest ranked agricultural economics journal since 2008. In 2010 the Agricultural and Applied Economics Association, formerly known as the American Agricultural Economics Association (same acronym, AAEA) decided to rebrand its second-ranked journal, and the Review of Agricultural Economics became Applied Economics Perspectives and Policy (AEPP). In terms of impact factor, the AEPP journal is now the second best in the field of agricultural economics after Food Policy, ahead of the leading AAEA journal, the American Journal of Agricultural Economics.

In short, (agricultural and food) policy analysis has become a best seller, and demand and supply of rigorous policy evaluations have grown very rapidly. From an era of paucity of quantitative evaluations, we have moved to abundance. Beyond societal interest, this trend has been driven by the amazing progress in data availability, and the evolution in user-friendly econometric software has been equally rapid.

As readily available data and software fertilize policy evaluation studies, the academic community needs to set higher methodological standards to defend the credibility and robustness of the findings. Without claiming the authority to define those standards, this manuscript has the objective to review the main quantitative methods currently employed in food policy evaluation, more specifically those targeting the causal identification of policies, and explicit the key assumptions they rest on. We restrict our range of applications to the analysis of policies targeting nutrition outcomes. There are not many comprehensive work on impact evaluation methods that are specific to nutrition policies (Babu *et al.*, 2016), and not many reviews of the policy evidence consider the credibility of causal inference methods (see e.g. Capacci *et al.*, 2012; Mazzocchi, 2017)

More specifically, the focus of this article is on the application of quasi-experimental methods when secondary data are used for ex-post assessment of food policies. While these "counterfactual" approaches are relatively young within this research field, they are rapidly becoming a minimum standard for causal inference in absence of randomization studies. The 2021 Nobel prize in economics has been awarded to David Card, Joshua Angrist, and Guido Imbens, three key contributors to methodological and empirical research on causal inference with observational data[1]. As it happens with most social science research objectives, economic policy analysis faces relevant challenges in drawing causal inference from randomized experiments[2]. Even in the less frequent situations where experimental evidence can be collected, the findings can be hardly generalized to be useful in other contexts. Thus, economists have historically relied on observational data in their evaluations of public policies, hence the need to address biases from the lack of randomization.

The article is structured as follows. We first discuss the opportunities and limitations in the data avail-

[1] See the document on the scientific background for the Nobel Prize, "Answering causal questions using observational data", https://www.nobelprize.org/uploads/2021/10/advanced-economicsciencesprize2021.pdf

[2] Still, the application of the experimental approach to economic problems has also generated important results. As one anonymous reviewer points out, the 2019 Nobel Prize was awarded to Esther Duflo, Abhijit Banerjee et Michael Kremer also in recognition of their application of the experimental approach "to alleviate global poverty".

able to researchers, especially in relation to the choice of adequate outcome variables to evaluate nutrition policies (Section 2). Then, we provide a short overview of the main quasi-experimental approaches to identify the causal effect of policies, with an emphasis on the assumptions they rest on, and whether and how they can be tested, as well as some approaches to demonstrate the robustness and validity of the causal findings (Section 3). Finally, we draw some take-home messages and suggest directions for future research.

2. DATA AND MEASUREMENT

What is the goal of nutrition policy? Such question is only apparently trivial, if one thinks what "improving nutrition" means. It is rather obvious that the ultimate aim of the policy is to improve human health, thus evaluations should rely on health outcomes. Unfortunately, the cause-effect path between improved nutrition and health outcomes is not immediate, and subject to major uncertainties. Hence, it is not surprising that most empirical evaluations of nutrition policies look at their short- to medium-term effects on intermediate outcomes, such as food choices or diet quality indicators, which in turn are health predictors[3]. The definition of these intermediate outcomes, however, is also subject to a variety of measurement-related issues.

Food choices not only vary across individuals, but also within individuals. Our Christmas food choices are likely to differ from those preceding the summer season, we may want to compensate on Mondays our week-end eating and drinking choices, and after a heavy lunch we may opt for a light dinner. Thus, a first question refers to the time interval which matters to define our baseline outcome indicator.

In nutrition science, the gold standard is the dietary record approach (Thompson and Byers 1994), the amount of food and beverages intake is recorded through a diary kept over a period of few days, normally no more than 7 consecutive days. This minimizes the memory bias, but may generate a fatigue effect (too much effort to keep the diary), and a behavioural bias associated with a "learning-by-doing" effect, as participants become aware of their eating patterns as they record them, and may alter their diets accordingly. An alternative approach rests on 24-hour recalls, which requires the respondent to recall and report all the food and beverages consumed during the previous day. While the task is not particularly burdensome and potentially more accurate, it fails to capture variation between days. This issue may be mitigated by appropriate sampling designs, assuming that heterogeneity across individuals belonging to a specific population group and interviewed at different times reflect – at least on aggregate – the average choices and intertemporal substitutions of individuals in the same group. A third nutrition-focused alternative is the Food Frequency Questionnaire (FFQ), which records the "usual" frequency of consumption of a list of food items. FFQs can be acceptable to measure average individual behaviours, but they are usually less accurate in quantifying intakes. Despite this, they are cheap and simple, and place a low burden for participants, which made them a commonly used dietary assessment tool (Thomson et al. 2003). Key food security indicators (e.g. the Food Consumption Score by the World Food Program) are based on FFQs.

Although this type of data has become relatively more common in food policy analysis, especially in development studies, economists remain concerned about the quality of measurement tools which depend on some form of self-assessment and have a component of social desirability bias (Grimm 2010). For example, Lissner (2002) shows that selective underreporting by obese individuals occurs with almost all methods of dietary assessments which rest on self-reports. Furthermore, nutrition survey data have a limited coverage of key food policy covariates, often failing to record the prices faced by individuals, their incomes, and consumption of non-food items.

This is why purchase data remain the preferred source of outcome indicators for economic studies, especially in the scanner data era. These large data sets not only allow to monitor individual daily transactions by individual household over several years, but they also have been augmented to provide detailed nutrient information at the level of unique product codes, as well as detailed data on purchase outlets, and household characteristics (Muth *et al.*, 2020; Biondi *et al.*, 2022). In household budget surveys, households record purchases through one- or two-week diaries, and data suffer by the aforementioned potential biases, although the lack of an explicit nutrition focus should mitigate social desirability biases. In consumer panels based on home scanners, participants scan universal product codes of all products taken home after each shopping trip. Point-of-sale scanner data are another rich source and provide measurements of sales volumes and prices, but cannot be related to individual consumer characteristics.

[3] When data allow to do so, causal mediation analysis is a powerful approach which supports the identification of causal chains, i.e. a causal estimate which goes beyond the total effect of the treatment on the outcome, and also identify the indirect effect that occurs due to one or more mediating variables. For a comprehensive overview, see Celli (2022).

Obviously, purchases are only a proxy of actual intakes, and the fact that these measures are at the household level is one serious shortcoming. Still, an underused opportunity of large scanner data-set is the possibility to monitor the transaction of one-member households over several years. While this is clearly a selected sample of the overall population, a time series of thousands of high-frequency data for the same individuals could be a unique setting for causal identification for policy evaluations.

To show the implications of using different outcome measures and ignoring self-reporting biases, we report in Table 2 some aggregate figures on attitudes towards fruit and vegetable consumption, self-assessed intakes and purchase-based intakes in the UK since the start of the 2003 national Five-a-Day campaign. The data come from nationally representative surveys and were not collected specifically to evaluate the campaign.

The first row captures the awareness impact of the information policy. During the first year of the policy, the average perception of what constitutes an optimal consumption was 4.4 portion per day. Over time, the campaign has been successful in increasing this target towards the "5-a-day" objective. The second row displays self-assessed intakes and is a clear demonstration of what social desirability means. While in 2003 participants were reporting an intake below the optimal target, in 2007 they were declaring an (average) consumption well above the ultimate policy objective. Unfortunately, when assessing intakes based on more objective purchase data, we find that the increase has been quite modest, and well below the perceived optimal intake. Clearly, the assessment of the policy effectiveness heavily depends on which outcome we choose to focus on.

There is no such thing as the perfect outcome variable and the quality of data is very heterogeneous. Rather than an excuse to discard quantitative policy evaluations, this should push researchers to discuss their data sources in great detail, acknowledge any limitation and adopt appropriate countermeasures and robustness checks.

Table 2. Knowledge, self-assessed consumption purchases of fruit and vegetables in the UK, number of portions per person per day (years 2003, 2006, 2007).

	2003	2006	2007
Optimal intake (reported)	4.4	4.6	4.8
Self-assessed intake	3.4	5.2	5.6
Assessed intake from purchases	3.7	4	3.9

Source: Our processing on data from UK Consumer Attitude Survey and Expenditure and Food Survey (various years).

A list of secondary data sources potentially relevant for nutrition policy evaluation is provided in Table 3. While individually these sources suffer from a variety of shortcomings, some may be addressed by adopting methods which integrate data from different sources, even if they do not pertain to the same subjects. Various techniques enable to combine two or more dataset. For example, Blundell *et al.* (2008) match consumer expenditure data from repeated cross-sectional consumer surveys with longitudinal data providing accurate information on incomes. Other techniques exist to combine information from different surveys (see the review in Lohr and Raghunathan, 2017). Furthermore, data collected as repeated cross-sections – as is the case for most of national household budget surveys – can be restructured into pseudo-panels by aggregating individual observations into homogeneous groups (e.g. same age group, same gender, same income bracket, etc.) which become the panel unit (Deaton 1985).

3. METHODS FOR CAUSAL INFERENCE

How do we know that it is rain that leads people to open their umbrella, and not open umbrellas that cause rain? If we had a spreadsheet showing (cross-sectional) data on the presence of rain and open umbrellas, statistics could definitely confirm that the two things are connected, and bring evidence that it is much less likely to find open umbrellas on sunny days. However, without some manipulation, statistics without prior theoretical knowledge is unable to infer causality from mere observational data. One way out in economics (and in life), is the assumption that what happens earlier is more likely to be the cause than the effect, but this reasonable simplification is often useless[4]. Suppose the government lowers VAT on healthy foods in year *t*, and in year *t+1* people consume less healthy foods. Using again our common sense and theoretical knowledge, we know that lowering VAT cannot cause lower consumption, but previous trends or other confounding factors (e.g. prices going up) are messing up with our attempt at causal inference.

[4] Indeed, estimates from correctly specified structural models drawing from validated economic theories can return good causal estimates. Once we know that rain causes open umbrella, and we have enough information to correctly specify our model (e.g. weekday, time of the day, ecc.), we can estimate the relationship between the amount of rain and the density of umbrella, and check that our estimator meets the desired economic properties. Our focus is in the (frequent) situation where theory provides insufficient guidance, or lack of information leads to biased estimates. As discussed later in the article, quasi-experimental methods can be a powerful complement to structural models. We are grateful to an anonymous reviewer for soliciting this clarification.

Table 3. Secondary data sources relevant for nutrition policy evaluation.

Type of survey	Description
Nutrition surveys	Specifically aimed at monitoring people diet, usually through dietary records, recall or FFQ. They usually collect data on key individual characteristics, mostly demographics, but sometimes also on health status and attitudes.
Health surveys	Based on interviewing/questionnaires and objective measurements (e.g. blood tests, urine samples, etc.), health surveys record information on people subjective and objective health status. Other information often collected: health related behavior, demographic characteristics, lifestyle topics such as smoking habits or dietary habits.
Household budget surveys	Their scope is to collect information on household purchases over a period of one or two weeks, based on expenditure diaries and face-to-face interviews. They normally include detailed information on food purchases (at the food item level) and demographic information on the household. In most countries they are run every year. When purchased quantities are provided along with expenditures, it is possible to estimate average prices.
Scanner data	This type of data records expenditure, paid prices and purchased quantities at the most detailed product level (brand and pack size). Data are collected either at the point-of-sale through cash registers (retail panels) or at home by household panels equipped with a barcode scanning device (consumer panels). These data are collected by private companies, and in some cases combined with product label information, including nutrition information. Households may remain in consumer panels for several years, allowing for longitudinal analyses.
Opinion/omnibus/ attitude surveys	These surveys collect information for multiple purposes, often including measurement of opinions, beliefs, self-reported habits, attitudes, stated preferences, perceived health, lifestyle factors, etc. They may contain self-reported information about eating behaviours and knowledge, and sometimes anthropometric measures.
Food composition databases	Food composition tables contain information on the average nutrient content of raw and processed food items available in one country. They are useful in combination with other data-sets to associate food items with their nutrient content.
Audience measurement data	These data are normally used to monitor media consumption (radio, television, newspaper, magazines, websites, social networks). They are conducted by private market research companies, and when combined with purchase data can be useful to explore the exposure to advertising and the impact of advertising regulations.
Epidemiological studies	They provide information on the prevalence and incidence of diseases, morbidity, mortality and related risk factors. They are useful to predict and simulate the ultimate health outcome of a policy based on the estimated impact. They are population-specific and are very heterogeneous in terms of sample sizes and duration.
Administrative data	These data are collected for administrative purposes, but some may be useful for evaluation, e.g. population registers of births, deaths, tax records as well as information on household access to subsidies and financial support. Administrative sources may also help quantifying the policy cost.

This short account does not do justice to centuries of questions about how science should look at cause-effect relationships, since Francis Bacon, and the logic provided by John Stuart Mills in 1843 to frame scientific experiments. But the key elements that matter to our treatment are randomization and the potential outcome framework, first formalized by Fisher in 1925 and Neyman in 1923, respectively (see Boring, 1954). Interestingly, while these two essential elements behind randomized experimental designs have developed almost simultaneously, their combination to support causal inference with observational data took another half century, until the key contribution of Donald Rubin (see Imbens and Rubin, 2015). The rationale behind randomized assignment was that "the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated'' (Fisher, 1935, p. 19). In other words, randomization as an insurance against confounding factors. Meanwhile, Neyman had introduced the concept of potential outcomes: "Let *x* denote a possible outcome of the experiment consisting of drawing one ball from the *i*-th urn" (Splawa-Neyman *et al.*, 1990). Basically, before the draw, an undrawn ball can take any number, just like the health status of Schrödinger's cat is unknown before opening the box.

How these philosophical wanderings matter to policy evaluation becomes clearer when one considers the "fundamental problem of causal inference", also referred to as the Neyman-Rubin causal model (Holland 1986). Before any scientific (randomized) or natural (non-randomized) experiments, subjects may expect one of two "future" outcomes, either under treatment or in a non-treatment ("control") situation. For example, before the government approved the budget law in October 2021, an Italian shopper could envisage for the first week of January 2022 one level of sugar-sweetened beverage (SSB) purchases "under the soda tax", and another level of SSB purchases "without the soda tax". For the policy analyst, the perfect evaluation would require to observe both outcomes on the same subject. The outcome difference would be the impact of the SSB tax on that specific individual consumer, and repeating the analysis on many shoppers would return the full distribution of impacts. In absence of parallel words, we have to settle with observing a single outcome. The Italian government decided to postpone the introduction of the soda tax for the second consecutive year, and in January 2022 only the "no tax" outcome was observable.

3.1 Randomized experiments

In experiments, a Fisher-style randomized study is the solution to this fundamental problem. The difference between average outcomes from two random samples drawn from the same population, where only one of the two sample is treated, returns an average effect of the treatment. As a matter of facts, in absence of the treatment, two random samples from the same population return average outcomes from the same outcome distribution, and there is no reason why the difference in average outcomes should be significantly different from zero.

The above trick works very well with scientific experiments, but several complications emerge when the subjects of the experiment are humans. Even in medical randomized controlled trials, external validity of treatment effect estimates is all but granted. Designs of experiment for social and policy evaluation studies are even harder to be set up in a meaningful way, for example ensure real randomization, avoid compliance issues, control for a multitude of confounding factors that may act differently between the two groups during the experiment.

One key dimension to be considered is what sort of randomization drives the experiment. Random assignment to the treatment or control group is the prerequisite of randomized controlled trials. However, this only ensures that the two samples come from the same population, which is not necessarily the actual population of interest and may be self-selected, especially if participation is voluntary. It is hard to think about ethically acceptable trials where participation is compulsory. Thus, even a perfect RCT may return an estimate of the treatment effect which is affected by a selection bias, when the overall sample of participants is not representative of the target population.

An excellent review of the potential limitations of RCTs – especially for causal inference in economics – is provided in Deaton and Cartwright (2018). Randomized (food) policy experiments are quite rare. Some notable exceptions are the US Healthy Incentives program (Olsho et al. 2016), or the income support Progresa program in Mexico, where the government – not having enough budget to target all low-income families – randomized the villages where the policy was implemented (Gertler 2004).

Meeting all conditions that make a randomized experiment on food policy able to deliver a reliable estimate of the treatment effect is not a trivial task. Thus, it should not be maintained (as it is often the case in public health studies) that randomized experiments are the gold standard, and observational studies are a second-best options to learn about policy effectiveness. However, randomization might be the best route (and possibly the only one) when the objective is not the ex-post quantification of the policy impact, but rather an ex-ante assessment or the ranking of alternative policy instruments addressing the same policy objective. Even in cases where the estimate of overall effect sizes cannot be fully trusted, it is possible that the ranking of policy instruments in terms of their cost-effectiveness has an acceptable external validity. See, for example, the randomized experiment on food/voucher/cash transfer in Northern Ecuador (Hidrobo et al. 2014).

3.2 Quasi-experimental methods

We now go back to the goal of this article, and discuss how observational non-randomized data can provide ex-post evidence on the impact of a policy. To introduce this class of methods, it may be useful to recall some very standard and light notation.

Let y_{it} be the potential outcome for unit *i* if exposed to the policy, and y_{ic} the potential outcome for the same unit when not exposed to the policy. Suppose that in our target population some units are eventually "treated" and exposed to the policy and other units are not, but assignment to treatment is not necessarily random. This situation where the treated and control group are not the consequence of an explicit randomized design is commonly referred to as a "natural" experiment[5]. For

[5] There is no consistent definition of "natural experiment" in the literature, beyond the common consensus on the lack of explicit randomization. Some restrict the definition to those situations where randomization occurs "naturally", i.e. assignment to treatment is "as if random", even without the explicit intervention by the researcher. For our discussion, we consider a broader case where the assignment mechanism is unknown and unknowable by the researcher, but there is some external event which allows to regard such mechanism as probabilistic (for a detailed discussion see Titiniuk, 2019).

example, one may compare soft-drink consumption in a country exposed to a SSB tax (France) with a neighbouring country without the tax (Italy), or school fruit schemes where participation of schools to the program is voluntary. We use a binary indicator D_i to capture exposure to the policy, where D_i=1 when units are treated and 0 otherwise. At this stage, we consider a situation where we only have a cross-section of units observed in a single time period after the policy implementation, but we can extend later the notation to consider methods that rely on multiple time periods.

Consider the following identity, where the left-hand term is the average outcome difference between the treated and the control group:

$$E(y_{it}|D_i=1)-E(y_{ic}|D_i=0)=[E(y_{it}|D_i=1)-E(y_{ic}|D_i=1)]+ [E(y_{ic}|D_i=1)-E(y_{ic}|D_i=0)] \qquad (1)$$

This equation decomposes the difference in means in two parts, the actual average treatment effect on the treated population (ATT) and the selection bias (SB). On the right-hand side of the equation, the first square bracket $[E(y_{it}|D_i=1)-E(y_{ic}|D_i=1)]$ is the ATT, since it compares the average potential outcome under treatment and the average potential control outcome for the same population of individuals, those that are actually treated. Thus, the ATT is the objective of the evaluation, and indicates how much the policy changes the outcome of those that have been exposed to the policy.

The second square bracket is the selection bias $[E(y_{ic}|D_i=1)-E(y_{ic}|D_i=0)]$, which is the difference in the average potential control (i.e. without policy) outcome between the treated population and the control population. Under perfect randomization, there would be no reason for this difference to be significantly different from zero, as the two samples (treated and controls) would be extracted from the same population. Here, however, we deal with observational data. It is hard to think that even without the French SSB tax, France would report the same average soft drink consumption level as Italy. Thus, in order to get the ATT it is necessary to purge the outcome difference from our data from a "baseline difference", intended as the difference between the two groups in absence of the policy.

In order to estimate the ATT, a counterfactual estimate is necessary. One way is to try and estimate $E(y_{ic}|D_i=1)$, which is the outcome we would have observed on the treated group had the policy not been implemented. This would allow to obtain the ATT directly from the left-hand side of (1). A symmetric route is to try and estimate the SB, and the same counterfactual estimate is needed for this purpose.

A first operational step in that direction is the identification of the drivers of the SB. Why are the outcomes in the two groups expected to be different in the two group in absence of the policy? Why is the French consumption of soft drinks expected to be different from the Italian one, even without the SSB tax? We can start by listing those characteristics – other than the tax – that influence soft drink consumption, the many "confounding factors" which are balanced between the two groups when a randomized assignment is possible. The list is long, prices (of soft drinks and substitutes), incomes, levels of advertising, culture and tastes, temperatures and seasonality… Having good information on all potential confounders is a very lucky situation, possibly unreal. In many policy situations where subjects may self-select into treatment, namely in voluntary schemes, psychological drivers can play a major role and they are hardly measured in secondary surveys. Thus, we complete our notation by defining a vector of subject characteristics **x**, which is composed by a set of observed variables (or *observables*) $\mathbf{x_O}$ and a set of unobserved variables (*unobservables*) $\mathbf{x_U}$. Whether a variable ends up in the former or latter set depends on the contents of our dataset.

In a randomized setting, the policy impact could be obtained by a very simple regression model, corresponding to a mean comparison test:

$$y_i=\alpha+\beta D_i+\varepsilon_i \qquad (2)$$

Since we have no reason to think that there are other differences than the policy between the two groups, β is a consistent estimate of the ATT, and the variance of the 0-mean random error captures the variability in outcomes. Randomization is expected to balance both $\mathbf{x_O}$ and $\mathbf{x_U}$ between the two groups, but the researcher may want to test how well it worked, and test for significant differences in $\mathbf{x_O}$. A successful randomization should ensure that none exists.

Without randomization, we ideally want to control for any confounding factors. Thus, the policy model for observational data becomes

$$y_i=\alpha+\beta D_i+\gamma\mathbf{x_O}+\delta\mathbf{x_U}+\varepsilon_i \qquad (3)$$

Which still returns a consistent estimate of the ATT through β. Unfortunately, we do not have information on $\mathbf{x_U}$, which leads to an omitted variable problem. Quasi-experimental methods try to sort out the issue.

3.2.1 Propensity score matching

The class of methods based on propensity score matching (PSM) has been popular in health sciences, but it is hardly useful for causal inference without combining it with other quasi-experimental methods[6]. The reason is simple, the key assumption behind PSM (called unconfoundedness) is that there are no variables in $\mathbf{x}_U$. Any variable which matters to the outcome and is unevenly distributed between those exposed to the policy and those not treated must be either known or highly correlated with a known variable. In other words, an effective matching requires full knowledge of the structural model determining outcomes, or full information about the selection process. In such especially desirable situation, even OLS estimates of the model in equation (3) would provide a consistent estimate, even more efficient than PSM provided that the linearity assumption holds and there are no heterogeneous treatment effects. Not only, but authoritative recent studies have emphasized that improper application of PSM could lead to the opposite (and highly undesirable) result of increasing unbalances in unobservables, and lead to larger biases (King and Nielsen 2019).

Nevertheless, PSM is widely used, probably because it is effective in reducing dimensionality, it is an intuitive and relatively easy to teach method, and statistical packages offer fast implementation algorithms. Without indulging in technical details that are much better described elsewhere (see Caliendo and Kopeinig, 2008), PSM aims at balancing the distribution – or at least the means – of observables between the treated and the control samples. It does so by working on the control sample, by dropping observations, or by applying weights. For example, an observation in the treatment group can be matched with a single observation in the control group, or with a weighted average of observations from the control group. How this matching is accomplished depends on the matching algorithm, and there are many variants: nearest neighbour, radius, kernel and stratification matching being those most commonly implemented. The idea is that rather than matching on the full set of variables $\mathbf{x}_O$, a synthetic function of these variables is used, the propensity score. A propensity score is the probability of a unit to end up in the treatment group given its observed characteristic $\mathbf{x}_O$, and can be easily estimated via a probit or logit model. Matching on the probabilities estimated through these models is easier and more feasible than attempting to match all individual characteristics.

The key assumption to exploit PSM for causal inference is unconfoundedness, which requires that no relevant unobservable variable exists. Can this assumption be tested? Not directly, but propensity scores are based on the estimation of a binary dependent variable model, and goodness-of-fit measures for that model, e.g. the Pseudo-R^2 or the rate of correct predictions, provide some feedback. Even if we find that most of the covariates are relevant (significant) in explaining the assignment-to-treatment process, low goodness-of-fit diagnostics signal that our observables are not enough, and the unconfoundeness assumption is not credible, unless one accepts that unexplained variability only depends on random factors, quite a strong requirement. More sophisticated testing strategies exist, as the Rosenbaum bounds or IV-based tests (see DiPrete and Gangl, 2004), but one should be wary of any PSM study that does not provide strong evidence that the unconfoundedness assumption is met, as ATT estimates may otherwise be affected by large biases.

Beyond this, PSM requires overlapping of the propensity scores ranges between the treatment and control group. In a non-random setting we are likely to find higher propensity scores in the target group, and some of them might be too high to find the right match in the control group. In that case, unmatchable observations are dropped from the target group, which means that the estimated ATT does not refer to the original treated sample, but to the reduced one. This might become a major limitation for the ATT estimate. Imagine that in a voluntary food assistance the poorest individuals are very likely to participate, hence have very high propensity scores, but they are not accounted in the ATT estimate because no adequate match is found. Then, the ATT will measure the impact of the policy on a population which excludes those who benefit the most.

Relative to other methods, PSM evaluations are less popular in nutrition policy analysis, but several applications can be found in the literature. Clark and Fox (2009) apply matching methods to investigate the impact of the US School Breakfast and National School Lunch Programs on vitamin, mineral and sodium intakes. The method seems to be more popular among development economists, for example Abebaw *et al.* (2010) use PSM to estimate the effects of a food security program in Northwestern Ethiopia.

3.2.2 Instrumental Variables

Provided that one or more "good" instruments are available, IV estimators of the ATT work on the same

[6] As pointed out by an anonymous reviewer, propensity scores estimates remain a useful tool to reduce dimensionality, and/or as a complement to other methods. Also, PSM has advantages when dealing with heterogeneous treatment effects.

data structure of PSM, and allow to control for selection effects driven by both observables – through direct inclusion in the estimation equation – and unobservables, the latter through instrumenting. We discuss later the fuzzy concept of "good instrument", and how authors, reviewers and journal editors tend to diverge in their opinion about the validity of instruments. The interpretation of IV models is straightforward, as it suffices to consider model (3). In absence of information on $\mathbf{x}_U$, we face the econometrics textbook problem of omitted variables, so that all coefficient estimates are biased and inconsistent. Under an economics viewpoint, a parallel interpretation is that the selection variable D_i is endogenous, as the probability of being exposed to the policy depends on the outcome level. For example, schools located in high income and education areas where fruit consumption is high, are more likely to participate in school fruit schemes.

Provided we have one or more adequate instruments $\mathbf{w}$ to instrument D_i, we can control for the selection bias and obtain consistent ATT estimates, at the cost of giving up some efficiency. Statistical packages routinely provide IV-2SLS estimators where the first stage regression is again a binary dependent variable model, a probit or a logit. Note that the structural policy model (3) still accounts for unbalances in observables , which enter directly the model as they are expected to influence the outcome. Instead, instruments should be variables that we would not use as direct explanatory variables for the outcome, and should be exogenous. If we have access to such type of variables, the first stage binary regression would be the same used to estimate propensity scores, with $\mathbf{x}_O$ as explanatory variables, plus the instruments $\mathbf{w}$ which do not belong to $\mathbf{x}_O$ and do not enter (3).

Since IV encompasses PSM[7] and accounts for selection on unobservables, why don't researchers just rely on IV estimation? The problem is likely to be a familiar one for the experienced reader. First, we struggle to find reasonable instruments in the dataset. Second, we struggle to convince reviewers that our instrument choice is a good one. Unfortunately, there is no definitive test on the validity of instruments that can convince all actors in the publication process. The issue is a Catch-22 one. In order to show that an instrument is exogenous, it must be independent from the residuals of the structural (second stage) equation. However, this test is theoretically impossible, as we only obtain unbiased estimates of the residuals when we have an exogenous instrument. The empirical solution is to use several instruments, leave one out, estimate the structural equation residuals through the other instruments, then check the correlation between the excluded instrument and the estimated residuals. One can then repeat the procedure leaving out a different instrument each time. While such a strategy may provide some support to the instrument validity claims, it is an empirical one, and it is still grounded on the assumption that the included instrument are exogenous and the residual estimates are unbiased. If many of our instruments are endogenous, the procedure is useless. Thus, we still need to be convinced and convince others that the instruments make sense under an economic perspective.

The other interesting element is the trade-off between consistency and efficiency. If the instrument are reasonable, exogenous, and obviously significant in the first stage equation, then we can place some trust in the consistency of the ATT estimate in the second stage equation. However, the ATT will have a larger standard error, as we rely on predictions of the D_i variable in the second stage, a sort of propensity score augmented by the instruments. How much larger the standard error depends again on the goodness-of-fit of the first stage probit or logit equation. This time, however, a poor fit does not lead to systematic biases, it just inflates the standard errors, and with large data-sets this is not usually a problem.

A list of instruments used in the food policy literature is beyond the scopes of this article, although it would be an interesting reading. For example, Hofferth and Curtin (2005) investigate the effect of school lunch programs on the BMI of students. Participation to the lunch programs is voluntary for schools, and students need to have specific characteristics to be eligible for a free meal. These policy elements are clearly a source of endogenous selection. Public school attendance is used as an instrument, as it does not affect BMI directly, but it is strongly associated with the school program participation, since public schools are more likely to be part of lunch programs.

An alternative strategy resting on the use of instruments is the control function approach. This approach involves a first stage to model the exposure to the program, and a second stage where the individual probability of exposure is included as an additional variable on the right-hand side of the outcome model, to correct for the selection bias. The Heckman two-step estimator is the most widely used control function approach. For example, Butler and Raymond (1996) explore the impact

[7] A caveat is necessary. Just like in OLS, unbiased estimation through a regression model still requires that the linear specification is appropriate and treatment effects are homogeneous, whereas PSM is more flexible. However, there is extensive research to extend IV to deal with heterogeneous treatment effects (see e.g. Klein, 2010 and references therein), and propensity score matching can be used in combination with IV estimates, which is why we refer to encompassing here.

of household participation in US Food Stamp program on nutrient intakes of the elderly, using a variety of instruments, including household assets and distance to a food stamp office.

3.2.3 Regression Discontinuity Designs

For some specific policies, eligibility depends on the threshold value for a single continuous variable. Typical examples are policies designed around an administrative eligibility criterion based on age or income thresholds to allow access to food assistance programs or other subsidies. When such a sharp classification exists and the variable is known, the division between target and control units is straightforward. As this variable is most likely to be a key determinant for the outcome of interest, this also implies that there is no overlapping and two sub-population are hardly comparable.

In these cases, restricting the analysis to those units that are just below or just above the threshold is a potential solution. With a very large sample, the researcher might have a sufficient number of observations even after restricting the data-set. For example, if a policy is targeting subjects aged below 30, and we have a large data set including individuals within 6 months from their 30th birthday, the resulting sample is relatively homogeneous in terms of age, and splitting the sample in two groups through the date of birth is similar to randomizing assignment, and one should not expect major selection biases. A mere mean comparison test between the average outcomes could be a quite good estimate of the ATT.

However, one major caveat accompanies this estimate of the treatment effect, which is certainly valid in the selected neighbourhood of the cut-off point, but not necessarily for data points further away. In our example, we may get good and reliable estimates of the ATT for those aged 30, but we can say little about the policy effects on those that are aged 20 or 25 relative to those aged 35 or 40. Thus, ATTs estimated through RDD are characterized by limited external validity.

Furthermore, this threshold analysis commonly runs into two major issues: (1) the number of available observations around the cut-off value is not large; (2) the cut-off point may be associated with a number of confounding events creating discontinuities. For example, if the age cut-off is also the retirement age (e.g. 65), one may think that such an event creates relevant disparities between the target and control groups in variables that may in turn affect the outcome variable.

The first problem is addressed by relying on the functional relationship between the outcome and the assignment (running) variable. When such function is identifiable, it can be exploited to expand the sample of interest. To do that, we need to assume continuity, which means that without the policy the outcome would just follow the identified functional relationship with the running variable. The most basic functional form is a simple bivariate linear regression, and the policy impact would be captured by a sharp shift in the intercept as the running variable reaches the cut-off point. By exploiting this linear relationship, one is able to expand the sample and consider units that are further away from the threshold. This brings in a second assumption, linearity, which requires that the linear relationship is valid within the expanded neighbourhood of the cut-off point. Although few relationships between the outcome and the running variable are indeed linear, when the neighbourhood under consideration is still relatively small, then the linear approximation performs well and the ATT estimate becomes more credible (and efficient) for the sample of interest. In other words, its internal validity is higher. Clearly this introduces a trade-off between internal validity and efficiency. If we consider a large neighbourhood, we have more observations and a more efficient estimate of the ATT. However, observations become more heterogeneous, the linearity assumption becomes more influential, and there is less internal validity.

RDD deals well with unobservables when these are unlikely to differ substantially between the two groups within a small neighbourhood of the cut-off point. However, the crucial continuity assumption implies that there are no other major "jumps" in relevant outcome determinants at the same cut-off point. There are cases when this assumption is clearly challenged, for example when the cut-off value is one with administrative and legal relevance. For example, age cut-offs at 18 and 65 are common to several economic and health policy measures, or some income eligibility threshold levels can be similar across different policies in the same country, which complicates the attribution of the causal effect to a specific policy. In such situations, the only viable solution seems to be the inclusion in the model of covariates which help to control the confounding effects (Frölich and Huber 2019). More generally, one should test whether the continuity assumption holds simply by applying the same RDD model to relevant confounding factors, and the expectation is not to find significant discontinuities. The continuity assumption fails to hold when subjects have some control on the assignment variable. For example, one might delay some revenue (job offer) to maintain eligibility for a program based on income thresholds. If these behaviours ("bunching") are possible, then the continuity assumption is challenged and RDD becomes less credible[8].

Since the estimation of causal effects through RDD depends on assumptions on the neighbourhood size and the shape of the relationship between the outcome and the running variable, a number of extensions and variants in the estimation procedures exist. First, the optimal size of the window around the cut-off point (the bandwidth) may be also an output of the estimation algorithm. Second, non-parametric regressions allow to relax the assumption of a linear relationship, and place different weights on observations depending on how far they are from the cut-off point. Third, when the running variable does not determine a sharp cut-off (i.e. all individuals meeting the rule are treated), but only creates a shift in the probability to be treated, then fuzzy RDD better serves for the purpose. This is the case of voluntary policies, where not all eligible individuals are exposed, and/or when there are exceptions allowing participation of subjects that do not meet the cut-off eligibility requirement.

Including covariates, changing the bandwidth, allowing for non-linear relationships, or opting for a fuzzy design are all choices that may potentially lead to different results, which is why convincing robustness checks are not an optional feature in RDD studies. On the one hand, one may want to show that the estimate of the causal effect is relatively consistent across different choices. On the other hand, falsification tests add credibility to the identification strategy. For example, one may want to show that different cut-off points other than the one relevant to the analysis are not associated with discontinuities.

Although the range of policies that are suitable to this method is limited, and the aforementioned external validity caveat applies, RDD is considered a relatively powerful causal identification method. Sometime researchers have expanded the scope of RDD by considering time as the assignment variable with panel or time series data (see e.g. Aguilar *et al.*, 2021). In these exercises, the idea is that comparing outcome just before and after the time of the policy implementation, while exploiting some outcome-time relationship, may lead to the identification of the policy causal effect. However, this also leads to major differences in the requirements for successful identification relative to the standard RDD method, an issue which deserves careful consideration before one chooses "time" RDD over simpler event study models (Hausman and Rapson 2018).

Examples of RDD application to nutrition policies include the income-eligibility rule for the US School Lunch Program (Schanzenbach 2009), the removal of vending machines from secondary schools in France (Capacci *et al.*, 2018), the impact on nutrition and well-being of a new refugee assistance program in Kenya (MacPherson and Sterck 2021), and the effects of micro-credit on children nutrition in China (You 2013).

3.2.4 Difference-in-differences

Difference-in-differences (DID) has clearly become the most popular and widely applied quasi-experimental method for investigating the causal effects of nutrition policies. The rationale of the method is well known, and it allows to control for selection biases driven by unobserved factors when data from natural experiments are available before and after the policy, provided the appropriate assumptions hold.

The DID approach follows from the extension of equation (1) to account for multiple time periods. Consider its most basic formulation with two time periods, one before (period 0) and one after the policy implementation (period 1). In period 1, by reworking equation (1), one may obtain the ATT by subtracting the SB from the difference in means:

$$ATT=[E(y^1_{it}|D_i=1)-E(y^1_{ic}|D_i=0)]-[E(y^1_{ic}|D_i=1)-E(y^1_{\mathrm{ic}}|D_i=0)] \tag{4}$$

Where the superscript of the outcome variable indicates the time period. Assuming that the selection bias does not change between period 0 and period 1, the pre-policy data can be exploited to estimate the selection bias and the ATT can be rewritten as:

$$ATT=[E(y^1_{it}|D_i=1)-E(y^1_{ic}|D_i=0)]-[E(y^0_{ic}|D_i=1)-E(y^0_{\mathrm{ic}}|D_i=0)] \tag{5}$$

Now all terms on the right hand-side of the equation are observable. Since there is no policy in period 0 we observe the control outcomes for both the treatment and the control groups. The assumption of constant selection bias is usually referred to as the parallel (or common) trend requirement, since it implies that in without the policy the outcomes evolve at the same pace over time, and could be represented graphically as two parallel lines. Such assumption can (and must) be tested when data are available for multiple periods before the policy implementation.

One nice feature of the DID model is that the ATT can be estimated through a standard regression model

[8] Interestingly, this opens the way to relevant behavioural evaluations and estimation which exploit the possibility to identify manipulation (see Kleven, 2016).

on the outcomes, which also allows to control for all observed covariates $\mathbf{x_O}$:

$$y_i=\alpha+\beta D_i+\gamma T_i+\delta P_i+\vartheta \mathbf{x_O}+\varepsilon_i \tag{6}$$

Where T_i is a binary variable which is 1 in time periods after the policy implementation and 0 otherwise, and $P_i=D_iT_i$ is another binary variable which is 1 when observation i belongs to the treatment group (D_i=1) and is observed after the policy implementation (T_i=1). The coefficient estimate δ is the ATT.

If equation (6) is based on repeated cross-sections over multiple time periods ($t=1,\ldots,K$), one could test the parallel trend assumption by allowing (conditional) outcomes to evolve linearly over time before the policy implementation:

$$y_{it}=\alpha+\beta D_{it}+\gamma t+\theta(D_{it}\times t)+\vartheta \mathbf{x}_{\mathbf{O}it}+\varepsilon_{it} \qquad \forall t\in\{T_{it}=0\} \tag{7}$$

When θ=0, there are no differential trends between the treated and control groups. When $\theta\neq0$ one might still estimate a DID model augmenting (6) to allow for divergent linear trends, but such an extension should be supported by credible graphical evidence of a linear evolution of the conditional outcomes.

More informative (and efficient) estimates are derived from panel data, where the same units are observed over multiple time periods. There are several advantages in the generalized DID model for panel data (a two-way fixed effects panel regression), as (a) the inclusion of cross-sectional fixed effects further controls for constant unobserved heterogeneity across units; (b) the inclusion of time fixed effects allows to control for non-linear heterogeneity across time periods, and extensions to control for differential linear or even non-linear trends are possible; (c) it is possible to allow for different levels of policy intensity (e.g. differential tax rates). Consider, for example, the following model:

$$y_{it}=\alpha_i+\mu_t+\Sigma^{N}_{r=1}\beta_r(R_{ir}\times t)+\delta P_{it}+\vartheta \mathbf{x}_{\mathbf{O}it}+u_{it} \tag{8}$$

where the subjects belong to N different groups which may exhibit different linear trends, for example there are N regions or states, and R_{ir}=1 when the subject belongs to region r and is 0 otherwise. In this model P_{it} is not necessarily a binary variable, for example it might be a continuous variable between 0 (no policy) and 1 (full policy implementation). In this case, δ estimates the effect of full implementation.

With adequate panel data, it is theoretically possible to allow for non-linear differential trends:

$$y_{it}=\alpha_i+\mu_t+(\tau_t\times D_{it})+\delta P_{it}+\vartheta \mathbf{x}_{\mathbf{O}it}+u_{it} \tag{9}$$

The above specification allows for differential time fixed effects between treated and control subjects, but identification becomes quite challenging, and even impossible if the policy is implemented at the same time for all treatment units. An alternative is to omit δP_{it} and explore the evolution of the differential time fixed effects $\tau_t\times D_{it}$ over time, expecting that they change abruptly relative to their previous pattern for time periods following an effective policy.

Whatever the specification of the DID model, a thorough exploration of pre-existing trends when panel data allow to do so is a necessary but not so trivial task. Pre-testing may be affected by low power, and conditioning on pre-existing trends may lead to biases. An interesting review of these issues and a survey of recent papers in leading economics journals is provided in Roth (2022). Another important note of caution is needed for the estimation of two-way fixed effects panel DID models when the policy effects are heterogeneous across groups or time periods, as causal estimates of average treatment effects may be misleading. Alternative estimators have been proposed to address the issue (see e.g. de Chaisemartin and D'Haultfœuille, 2020).

The growing availability of panel data, especially commercial consumer panels with a high level of geographical details, has generated an exponential growth of DID models applied to the evaluation of the impact of fiscal policies on nutrition outcomes. For example, there is a high number of studies on national, state-level or even city-level taxes on sugar-sweetened beverage (see the review in Cawley *et al.*, 2019, or the report by Griffith *et al.* 2019). Beyond taxation, the DID approach has been applied to a variety of nutrition policies, including nutritional label regulations (Variyam 2008), calorie labelling in restaurant menus (Vadiveloo, Dixon, and Elbel 2011), school-based policies (Bhattacharya *et al.*, 2006), targeted subsidies (Griffith *et al.*, 2018), food assistance programs (Rahman 2016), information campaigns (Asirvatham *et al.*, 2017) advertising regulations (Dhar and Baylis 2011). The latter reference contains an example of how DID can be reinterpreted in applications lacking the time dimension, and even extended to the situation where multiple control groups can be considered. In Dhar and Baylis (2011) the impact on fast-food purchases of an advertising ban to children programs applied to TV channels in French-speaking Ontario is estimated through a Triple DID model using post-policy data only. The identification strategy rests on different target-control classifications, as both household without children and household with children in the near English-speaking Ontario region constitute potential control groups for household with children in Quebec, the target group.

3.2.5 Strategies based on structural models

An alternative approach is needed in situations where there is no natural counterfactual, for instance when a policy potentially acts on the whole population, as in a nationwide a public information campaign. As information policies may be expected to generate behavioural effects beyond the mere change of the average outcome, an option is to generate model-based counterfactual estimates. This approach is especially interesting when the behaviour of interest is well captured by a consolidated economic specification, and it is conveniently applicable when the pre-policy and post-policy data come from different (repeated) cross-sectional samples from the same population[9]. One may then express the outcome as the function of its determinants in each period:

$$y^0_i = f^0(\mathbf{x}^0_{i\mathrm{O}}) + \varepsilon^0_i$$

and

$$y^1_i = f^1(\mathbf{x}^1_{i\mathrm{O}}) + \varepsilon^1_i$$

The functions f^0 and f^1 have the same structural specification, but are characterized by different parameters. For example, f might be a demand function and the parameters represent price and income elasticities. As implied by the Lucas critique, a policy is likely to go beyond changing the average level of consumption, and also lead to a change in elasticities, hence the change from f^0 to f^1.

If the policy has no direct impact on the covariates $\mathbf{x}_{i\mathrm{O}}$, then the two set of estimates allow to evaluate the counterfactual outcome, which is estimated as $y^1_i = f^0(\mathbf{x}^1_{i\mathrm{O}})$. In in our example this is the level of consumption that would have been observed in period 1 had the population maintained the preference structure of period 0. The ATT is $f^1(\mathbf{x}^1_{i\mathrm{O}})\text{-}f^0(\mathbf{x}^1_{i\mathrm{O}})$. The approach can be modified to include constraints on behavioural parameters, for example one might require that some of them remain constant between the two time periods. Also, if there are variables in $\mathbf{x}^1_{i\mathrm{O}}$ that are significantly affected by the policy, and it is possible to disentangle such effect (e.g. an estimate of the change in public advertising expenditure, or of the price change associated with a tax), one might estimate the counterfactual through $f^0(\mathbf{x}^1_{i\mathrm{O}})$ where the relevant variables in $\mathbf{x}^1_{i\mathrm{O}}$ are purged from the policy effect.

When data are organized as panels or relatively long time series, alternative approaches based on structural models may rely on switching and time-varying parameter regressions, intervention or event study analyses. All of these models allow one or more parameters to change in response to the policy. The most basic formulation aims at estimating a sharp step (i.e. an intercept shift as in event studies) at the time of the policy implementation[10]. When data allow to do so, any parameter in the structural model can potentially change and evolve, either with a pre-determined shape (as in intervention analysis or switching regression) or through random shocks (as in time-varying parameters models).

An example of nutrition policy evaluation where the counterfactual is based on a structural model is provided in Capacci and Mazzocchi (2011), who explore the effects of the 5-a-day information campaign in the UK through a demand system. Attanasio *et al.* (2012) exploit randomisation in the Mexican program Progresa to discuss how structural models can improve program evaluations even in cases where evidence from experiments is available. Kim *et al.* (2001) estimate a switching regression model to capture the effect of the Nutrition Labelling and Education Act on diet quality in the US.

4. CONCLUDING REMARKS, EXTENSIONS AND PERSPECTIVES

This article aims to provide a critical overview of the current state-of-the-art in the field of nutrition policy evaluation using quasi-experimental data. It is not comprehensive in terms of the range of counterfactual methods potentially available to researchers, and by the time it will be published and read it might even be "not-so-current". However, until the recent past, nutrition policy evaluation has relied on a much more outdated toolbox relative to other fields, especially compared to labour and health policy analyses.

A ranking or a direct comparison of the different methods would not be a wise exercise, as the choice and credibility of quasi-experimental methods is heavily dependent on the plausibility of the underlying assumptions, and the quality and detail of the available data. Inevitably, empirical diagnostics on the quality of "counterfactual" causal inference must depend on observed

[9] Furthermore, structural models and theoretical knowledge are always a valuable complement in the estimation of regression-based quasi-experimental models, as the DID, RDD and IV approaches can follow a structural specification.

[10] The analogy with regression discontinuity design is considered in Section 3.2.3.

variables, which can be outcomes, covariates, or instruments. Still, the crucial assumptions refer to unobserved and unobservable variables, and in most cases no conclusive test is available, as discussed in this article.

Despite this necessary disclaimer, we believe that consolidated and emerging methods will need to deal with the key elements we emphasize. The central and obvious one is that the success of any causal identification depends on the validity of its underlying assumptions. Under a technical point of view, this requires validation of findings through appropriate tests for these assumptions – even when they are only suggestive and not conclusive -, together with robustness and falsification checks, and comparisons with alternative identification strategies and possibly even different data.

There are several variants of quasi-experimental methods that may improve causal inference. For example, when pre-policy data cover multiple periods and multiple non-treated groups (e.g. regions), the synthetic control method (SCM) is a popular option (Abadie *et al.*, 2015). Consider a situation where only one region is treated, and there are *n* non-treated regions. The principle is relatively straightforward, instead of using the *n* controls separately, they are artificially combined into a single control group as a weighted average. The weights are obtained through an optimization algorithm which minimizes – in each time period before the policy – the distance between the outcomes and the observed covariates measured in the target group and those obtained as the weighted average of the *n* values measured in the multiple control groups. In other words, the SCM allows not only to ensure the common trend between the treated region and the artificial control group, but also balances the covariates. Then, the weights can be applied in the post-policy period to obtain the counterfactual outcomes.

Other extensions allow to provide better insights on the impact of a policy by going beyond the average effect and considering characteristics of the ATT distribution. For example, the difference-in-difference method can be implemented through (panel) quantile regressions, as in the study on the effect of the India public distribution system on nutrient intakes (Chakrabarti *et al.*, 2018). Recent developments exploit evaluation techniques based on machine learning methods and LASSO estimators (Belloni *et al.*, 2017).

The growing availability of micro-data has brought more emphasis on the identification of heterogenous policy impacts, which poses serious challenges to the interpretation of the average treatment effect (whether ATE or ATT), to the point that in some cases it is not possible to estimate credible average effects. One of these situations is the potential non-compliance to the policy measure by treated subjects, as non-compliers become a third selected group whose members may be systematically different from both treated-compliers and non-treated subjects. A typical example is a policy where compliance is correlated with the treatment effect, for example adherence to nutrition guidelines is likely to depend on the distance between the current diet and the recommended one. One solution might be simply to ignore the compliance issue, consider all those exposed to the policy as the target group, then apply the appropriate method. Hence, the resulting estimate will not reflect the actual effect of the treatment, but rather to the average impact on those that are exposed even if they do not “take” the treatment, which is referred to as the average intention-to-treat (ITT) effect. Alternatively, one may want to consider only those subjects that are exposed and comply with the policy, obviously controlling for the additional selection effect between compliers and non-compliers. The latter approach aims to estimate local average treatment-effects (LATE), generally through an IV estimator (Imbens and Wooldridge 2009). More generally, it is not infrequent that the impact of a program varies across subjects just because of the nature of the intervention, for example personalized nutrition actions, hence effectiveness depends on individual subject characteristics. The recent methodological developments are directed at tackling this challenge and capture heterogeneous effects across population subgroups, typically by letting the treatment effect depend on subject characteristics. The developments in data availability and machine learning techniques are especially important to address treatment effects heterogeneity (Athey and Imbens 2017).

Under a broader economic evaluation perspective, as researchers we unfortunately face a trade-off between the econometric rigour of the identification strategy and the policy relevance of the findings. A typical example is the focus on immediate (and easier to measure) outcomes which may be distant from the ultimate goal of the policy. Do sugar taxes work? Typing this question into Google Scholar returns a little less than 300,000 references at the date we are writing, but we challenge readers to find studies with robust causal inference about their effect on morbidity or mortality. This does not mean that “reasonable” assessments and simulations of the health impact of sugar taxes do not exist, nor that the scarcity of ATT estimates for health outcomes depends on gaps in the quantitative evaluation toolbox. Obviously, the problem lies in the lack of adequate data. The desired effects of many nutrition policies only emerge in the medium-to-long term, and would require prospective cohort studies following people from the cradle to the grave. To the best of our knowledge, no

European country is running nutrition studies of this type, and probably the only good example is the cohort study which monitors cardiovascular disease, diet, physical activity on the population of Framingham in Massachusetts since 1948, now on the third generation of participants (Andersson *et al*., 2019). Public investments on more broadly representative and durable prospective cohort studies would generate more knowledge on policy effects than what a century of studies on causal inference has allowed us to do.

Under this perspective, the "big data" challenge for causal inference, the hot topic in data-rich environments, is less urgent for nutrition policy analysts. Instead, we believe that another big methodological challenge of the coming years will become especially relevant to nutrition policy, i.e. the ability to make adequate causal inference from observational data when multiple policies coexist over the data support window. The international history of policy failures in trying to improve diets and reduce obesity, together with a lack of conclusive evidence on longer term outcomes, has favoured the adoption of a "trial and error" policy approach, with a variety of overlapping policies. The coexistence of multiple interventions is clearly an obstacle for the causal inference approaches discussed in this review.

On the other hand, a key contribution to nutrition policy evaluation from economists and researchers is related to improving the specification of structural (behavioural) models of food choice. Just to mention the most apparent challenge, few evaluation studies succeed in properly modelling dynamic behaviours when using secondary data, which is a requirement to consider habits, intertemporal compensations, discounting, stockpiling. In our discussion, we have underlined that structural models and the proper consideration of prior theoretical knowledge make them an ideal complement rather than an alternative to quasi-experimental methods.

Finally, causal inference techniques might bring major benefits to the policy evidence-base when combined with other decision support tools that are becoming increasingly popular in nutrition, stochastic microsimulation methods (see e.g. Emmert-Fees *et al*., 2021). Robust evidence on proximal outcomes from quasi-experimental methods could be valued in combination with simulation methods able to account for longer term effects, dynamic behaviours and heterogeneous impacts.

ACKNOWLEDGEMENTS

This work was funded by a grant from Italy's Ministry of University and Research as part of the Joint Programming Initiative (JPI) 'A Healthy Diet for a Healthy Life', a research and innovation initiative of EU member states and associated countries. The funder had no role in the design or writing of this article. We are grateful to the Editors and two anonymous reviewers for their feedback and suggestions.

REFERENCES

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. 'Comparative Politics and the Synthetic Control Method'. *American Journal of Political Science* 59 (2): 495–510. https://doi.org/10.1111/AJPS.12116.

Abebaw, Degnet, Yibeltal Fentie, and Belay Kassa. 2010. 'The Impact of a Food Security Program on Household Food Consumption in Northwestern Ethiopia: A Matching Estimator Approach'. *Food Policy* 35 (4): 286–93. https://doi.org/10.1016/J.FOODPOL.2010.01.002.

Aguilar, Arturo, Emilio Gutierrez, and Enrique Seira. 2021. 'The Effectiveness of Sin Food Taxes: Evidence from Mexico'. *Journal of Health Economics* 77 (May): 102455. https://doi.org/10.1016/J.JHEALECO.2021.102455.

Andersson, Charlotte, Andrew D. Johnson, Emelia J. Benjamin, Daniel Levy, and Ramachandran S. Vasan. 2019. '70-Year Legacy of the Framingham Heart Study'. *Nature Reviews Cardiology 2019 16:11* 16 (11): 687–98. https://doi.org/10.1038/s41569-019-0202-5.

Asirvatham, Jebaraj, Paul E McNamara, and Kathy Baylis. 2017. 'Informational Campaign Effects of the Nutrition Labeling and Education Act (NLEA) of 1990 on Diet'. *Cogent Social Sciences* 3 (1): 1327684. https://doi.org/10.1080/23311886.2017.1327684.

Athey, Susan, and Guido W. Imbens 2017. 'The State of Applied Econometrics: Causality and Policy Evaluation', *Journal of Economic Perspectives*, 31(2), pp. 3–32. doi: 10.1257/JEP.31.2.3.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago. 2012. 'Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA'. *Review of Economic Studies* 79 (1): 37–66. https://doi.org/10.1093/restud/rdr015.

Babu, Suresh, Shailendra Gajanan, and J Arne Hallam. 2016. *Nutrition Economics: Principles and Policy Applications*. Academic Press.

Belloni, Alexandre, Chernozhukov, Victor, Fernández-Val, Ivan, and Christian Hansen 2017. 'Program Evaluation and Causal Inference With High-Dimensional Data'. *Econometrica* 85 (1): 233–98. https://doi.org/10.3982/ecta12723.

Bhattacharya, Jayanta, Janet Currie, and Steven J. Haider. 2006. 'Breakfast of Champions?' *Journal of Human Resources* XLI (3): 445–66. https://doi.org/10.3368/JHR.XLI.3.445.

Biondi, Beatrice, Sara Capacci, and Mario Mazzocchi. 2022. 'Discrete Choice Models and Continuous Demand Systems in the Scanner Data Age'. In *A Modern Guide to Food Economics*, edited by Jutta Roosen and J. E. Hobbs. Cheltenam (UK): Edward Elgar Publishing Ltd.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. 'Consumption Inequality and Partial Insurance'. *American Economic Review* 98 (5): 1887–91. https://doi.org/10.1257/aer.98.5.1887.

Boring, Edwin G. 1954. 'The Nature and History of Experimental Control.' *The American Journal of Psychology* 67 (4): 573–89. https://doi.org/10.2307/1418483.

Butler, J. S., and Jennie E. Raymond. 1996. 'The Effect of the Food Stamp Program on Nutrient Intake'. *Economic Inquiry* 34 (4): 781–98. https://doi.org/10.1111/j.1465-7295.1996.tb01410.x.

Caliendo, Marco, and Sabine Kopeinig. 2008. 'Some Practical Guidance for the Implementation of Propensity Score Matching'. *Journal of Economic Surveys* 22 (1): 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x.

Capacci, Sara, and Mario Mazzocchi. 2011. 'Five-a-Day, a Price to Pay: An Evaluation of the UK Program Impact Accounting for Market Forces'. *Journal of Health Economics* 30 (1). https://doi.org/10.1016/j.jhealeco.2010.10.006.

Capacci, Sara, Mario Mazzocchi, Bhavani Shankar, José Brambila Macias, Wim Verbeke, Federico J.A. Pérez-Cueto, Agnieszka KoziolŁ-Kozakowska, et al. 2012. 'Policies to Promote Healthy Eating in Europe: A Structured Review of Policies and Their Effectiveness'. *Nutrition Reviews* 70 (3). https://doi.org/10.1111/j.1753-4887.2011.00442.x.

Capacci, Sara, Mario Mazzocchi, and Bhavani Shankar. 2018. 'Breaking Habits: The Effect of the French Vending Machine Ban on School Snacking and Sugar Intakes'. *Journal of Policy Analysis and Management* 37 (1): 88–111. https://doi.org/10.1002/pam.22032.

Cawley, John, Anne Marie Thow, Katherine Wen, and David Frisvold. 2019. 'The Economics of Taxes on Sugar-Sweetened Beverages: A Review of the Effects on Prices, Sales, Cross-Border Shopping, and Consumption'. https://doi.org/10.1146/annurev-nutr-082018.

Celli, Viviana (2022). Causal mediation analysis in economics: Objectives, assumptions, models. *Journal of Economic Surveys*, 36(1), 214-234.

Chakrabarti, Suman, Avinash Kishore, and Devesh Roy. 2018. 'Effectiveness of Food Subsidies in Raising Healthy Food Consumption: Public Distribution of Pulses in India'. *American Journal of Agricultural Economics* 100 (5): 1427–49. https://doi.org/10.1093/AJAE/AAY022.

Clark, Melissa A., and Mary Kay Fox. 2009. 'Nutritional Quality of the Diets of US Public School Children and the Role of the School Meal Programs'. *Journal of the American Dietetic Association* 109 (2): S44–56. https://doi.org/10.1016/J.JADA.2008.10.060.

de Chaisemartin, Clément, and Xavier D'Haultfœuille, X. 2020. 'Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects', *American Economic Review* 110(9), pp. 2964–96. doi: 10.1257/AER.20181169.

Deaton, Angus. 1985. 'Panel Data from Time Series of Cross-Sections'. *Journal of Econometrics* 30 (1–2): 109–26. https://doi.org/10.1016/0304-4076(85)90134-4.

Deaton, Angus, and Nancy Cartwright. 2018. 'Understanding and Misunderstanding Randomized Controlled Trials'. *Social Science and Medicine* 210 (August): 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005.

Dhar, Tirtha, and Kathy Baylis. 2011. 'Fast-Food Consumption and the Ban on Advertising Targeting Children: The Quebec Experience'. *Journal of Marketing Research* 48 (5): 799–813. https://doi.org/10.1509/jmkr.48.5.799.

DiPrete, Thomas A., and Markus Gangl. 2004. 'Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments'. *Sociological Methodology* 34: 271–310. https://doi.org/10.1111/j.0081-1750.2004.00154.x.

Emmert-Fees, Karl M.F., Florian M Karl, Peter Von Philipsborn, Eva A Rehfuess, and Michael Laxy. 2021. 'Simulation Modeling for the Economic Evaluation of Population-Based Dietary Policies: A Systematic Scoping Review'. *Advances in Nutrition* 12 (5): 1957–95. https://doi.org/10.1093/advances/nmab028.

Fisher, Ronald A. 1935. 'The Logic of Inductive Inference'. *Journal of the Royal Statistical Society* 98 (1): 39. https://doi.org/10.2307/2342435.

Frölich, Markus, and Martin Huber. 2019. 'Including Covariates in the Regression Discontinuity Design'. *Journal of Business and Economic Statistics* 37 (4): 736–48. https://doi.org/10.1080/07350015.2017.1421544.

Gertler, Paul. 2004. 'Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment'. In

American Economic Review, 94:336–41. https://doi.org/10.1257/0002828041302109.

Griffith, Rachel, Stephanie von Hinke, and Sarah Smith. 2018. 'Getting a Healthy Start: The Effectiveness of Targeted Benefits for Improving Dietary Choices'. *Journal of Health Economics* 58 (March): 176–87. https://doi.org/10.1016/j.jhealeco.2018.02.009.

Griffith, Rachel, Martin O'Connell, K. Smith, and R. Stroud. 2019. *The Evidence on the Effects of Soft Drink Taxes*. *The Institute for Fiscal Studies*. IFS Briefing Note BN255.

Grimm, Pamela. 2010. 'Social Desirability Bias'. In *Wiley International Encyclopedia of Marketing*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444316568.wiem02057.

Hausman, Catherine, and David S. Rapson. 2018. 'Regression Discontinuity in Time: Considerations for Empirical Applications'. *Https://Doi.Org/10.1146/Annurev-Resource-121517-033306* 10 (October): 533–52. https://doi.org/10.1146/ANNUREV-RESOURCE-121517-033306.

Hidrobo, Melissa, John Hoddinott, Amber Peterman, Amy Margolies, and Vanessa Moreira. 2014. 'Cash, Food, or Vouchers? Evidence from a Randomized Experiment in Northern Ecuador'. *Journal of Development Economics* 107 (March): 144–56. https://doi.org/10.1016/j.jdeveco.2013.11.009.

Hofferth, Sandra L, and Sally Curtin. 2005. 'Poverty, Food Programs, and Childhood Obesity'. *Journal of Policy Analysis and Management*. https://doi.org/10.1002/pam.20134.

Holland, Paul W. 1986. 'Statistics and Causal Inference'. *Source: Journal of the American Statistical Association* 81 (396): 945–60.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference: For Statistics, Social, and Biomedical Sciences an Introduction*. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. 'Recent Developments in the Econometrics of Program Evaluation.' *Journal of Economic Literature*, 47 (1): 5-86. https://doi.org /10.1257/jel.47.1.5

Kim, Sung Yong, Rodolfo M. Nayga, and Oral Capps. 2001. 'Food Label Use, Self-Selectivity, and Diet Quality'. *Journal of Consumer Affairs* 35 (2): 346–63. https://doi.org/10.1111/J.1745-6606.2001.TB00118.X/FORMAT/PDF.

King, Gary, and Richard Nielsen. 2019. 'Why Propensity Scores Should Not Be Used for Matching'. *Political Analysis* 27 (4): 435–54. https://doi.org/10.1017/pan.2019.11.

Klein, Tobias J. 2010 'Heterogeneous treatment effects: Instrumental variables without monotonicity?', Journal of Econometrics, 155(2), pp. 99–116. doi: 10.1016/J.JECONOM.2009.08.006.

Kleven, Henrik Jacobsen. 2016. 'Bunching'. *Annual Review of Economics*. Annual Reviews. https://doi.org/10.1146/annurev-economics-080315-015234.

Lissner, Lauren. 2002. 'Measuring Food Intake in Studies of Obesity'. *Public Health Nutrition* 5 (6a): 889–92. https://doi.org/10.1079/phn2002388.

Lohr, Sharon L., and Trivellore E. Raghunathan. 2017. 'Combining Survey Data with Other Data Sources'. *Https://Doi.Org/10.1214/16-STS584* 32 (2): 293–312. https://doi.org/10.1214/16-STS584.

MacPherson, Claire, and Olivier Sterck. 2021. 'Empowering Refugees through Cash and Agriculture: A Regression Discontinuity Design'. *Journal of Development Economics* 149 (March): 102614. https://doi.org/10.1016/j.jdeveco.2020.102614.

Mazzocchi, Mario. 2017. 'Ex-Post Evidence on the Effectiveness of Policies Targeted at Promoting Healthier Diets'. *FAO Trade Policy Technical Notes* 19 (November).

Muth, Mary K., Abigail M. Okrent, Chen Zhen, and Shawn A. Karns. 2020. *Using Scanner Data for Food Policy Research*. London: Elsevier. https://doi.org/10.1016/b978-0-12-814507-4.09993-4.

Olsho, Lauren E.W., Jacob A. Klerman, Parke E. Wilde, and Susan Bartlett. 2016. 'Financial Incentives Increase Fruit and Vegetable Intake among Supplemental Nutrition Assistance Program Participants: A Randomized Controlled Trial of the USDA Healthy Incentives Pilot'. *The American Journal of Clinical Nutrition* 104 (2): 423–35. https://doi.org/10.3945/AJCN.115.129320.

Rahman, Andaleeb. 2016. 'Universal Food Security Program and Nutritional Intake: Evidence from the Hunger Prone KBK Districts in Odisha'. *Food Policy* 63 (August): 73–86. https://doi.org/10.1016/j.foodpol.2016.07.003.

Roth, Jonathan. 2022, forthcoming. 'Pre-test with Caution: Event-Study Estimates after Testing for Parallel Trends'. *American Economic Review: Insights.* Available at: https://www.aeaweb.org/articles?id=10.1257/aeri.20210236

Schanzenbach, Diane Whitmore. 2009. 'Do School Lunches Contribute to Childhood Obesity?' *Journal of Human Resources* 44 (3): 684–709. https://doi.org/10.3368/JHR.44.3.684.

Splawa-Neyman, Jerzy, D M Dabrowska, and T P Speed. 1990. 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.' *Science* 5 (4): 465–72.

Thompson, Frances E., and Tim Byers. 1994. 'Dietary Assessment Resource Manual'. *Journal of Nutrition* 124 (11 SUPPL.). https://doi.org/10.1093/jn/124.suppl_11.2245s.

Thomson, Cynthia A., Anna Giuliano, Cheryl L. Rock, Cheryl K. Ritenbaugh, Shirley W. Flatt, Susan Faerber, Vicky Newman, et al. 2003. 'Measuring Dietary Change in a Diet Intervention Trial: Comparing Food Frequency Questionnaire and Dietary Recalls'. *American Journal of Epidemiology* 157 (8): 754–62. https://doi.org/10.1093/AJE/KWG025.

Titiunik, Rocio 2021, *Natural Experiments*. In: Green DP, Druckman JN, editors. Advances in Experimental Political Science. Cambridge: Cambridge University Press, 103-129.

Vadiveloo, Maya K., L. Beth Dixon, and Brian Elbel. 2011. 'Consumer Purchasing Patterns in Response to Calorie Labeling Legislation in New York City'. *International Journal of Behavioral Nutrition and Physical Activity* 8 (1): 1–9. https://doi.org/10.1186/1479-5868-8-51/TABLES/6.

Variyam, Jayachandran N. 2008. 'Do Nutrition Labels Improve Dietary Outcomes?' *Health Economics* 17 (6): 695–708. https://doi.org/10.1002/HEC.1287/FORMAT/PDF.

You, Jing. 2013. 'The Role of Microcredit in Older Children's Nutrition: Quasi-Experimental Evidence from Rural China'. *Food Policy* 43 (December): 167–79. https://doi.org/10.1016/J.FOODPOL.2013.09.005.