Expert Opinion

Statistical-methodological evaluation

of modeling studies

on the effects of a sugar tax

Dr. Katharina Schüller,

Accredited European Statistician (AEUStat)

Scientific Consultation:

Prof. Dr. Ralf Münnich

Prof. Dr. Danny Pfeffermann

Statistical-methodological evaluation of modeling studies on the effects of a sugar tax

Dr. Katharina Schüller, AEUStat[1]

Scientific Consultation:
Prof. Dr. Ralf Münnich[2], Prof. Dr. Danny Pfeffermann[3]

Research Assistants[4]: Simon Drauz, Sarah Huber, Lisa Kauck, Felix Lütge, Emma Morck, Maryna Puchkovska, Simon Rechberger, Lilian Schumacher, Dr. Chakresh Singh, Max Woldrich

Munich

April 2025

[1]Managing Director STAT-UP GmbH, Member of the Board of Directors of the German Statistical Society
[2]Professor Economic and Social Statistics at the University of Trier
[3]Professor of Social Statistics at the University of Southampton
[4]Staff STAT-UP GmbH

STAT-UP

Statistical Consulting & Data Science GmbH

Augustenstr. 5

D - 80333 München

Managing Director: Dr. Katharina Schüller, AEUStat

Telephone: +49 (89) 34 077 451

Fax: +49 (89) 34 077 453

E-Mail: info@stat-up.com

Web: www.stat-up.com

Executive summary

This report analyzes and evaluates a selection[5] of modeling studies on the effects of a sugar tax from a statistical and methodological perspective. Modeling studies are an important tool for assessing the (potential) impact of regulatory measures. They can be theoretically suitable for creating an evidence base for regulatory measures – but only if they are based on precise data and on robust and carefully validated assumptions.

> The report concludes that from a statistical and methodological perspective, none of the evaluated modeling studies on the effects of a sugar tax meet the high quality standards required for evidence-based policy decisions.

One of the main reasons behind this is that data of sufficient quality is not fully available. In the absence of sufficiently detailed data sets, numerous assumptions have to be made that do not always adequately reflect reality. This means that the results of the studies examined are subject to considerable uncertainty.

Key findings on modeling studies

Based on the existing research on modeling studies and the evaluations of the individual studies carried out in this report, the following key findings on the general methodological limitations of modeling studies can be derived:

1. Modeling studies that investigate (hypothetical) regulatory measures, e.g., the introduction of a sugar tax in Germany, are not independent evidence of causal relationships; rather, the existence of causal effects between influencing and target variables is assumed beforehand in the chosen modeling methods.

2. The existence and quantification of causal mechanisms is based on assumptions, e.g., the transferability of earlier empirical studies. Such assumptions are inevitably associated with uncertainties in the modeling results. Every decision for a certain assumption is at the same time a decision against alternative possibilities. Modeling results are therefore significantly influenced by the underlying assumptions.

3. The validity of modeling studies is heavily dependent on the availability and quality of the required empirical input data. Modeling studies usually obtain these from external sources and assume their representativeness and accuracy, often without explicitly checking either of them.

4. Inaccurate, biased and/or non-representative data, as well as incorrect or overly simplistic assumptions, lead to model results that may be consistent within the model, but do not accurately reflect reality.

[5] The selection includes the following studies:

- Emmert-Fees et al. (2023), Germany: Projected health and economic impacts of sugar-sweetened beverage taxation in Germany: A cross-validation modeling study.
- Schwendicke and Stolpe (2017), Germany: Taxing sugar-sweetened beverages: Impact on overweight and obesity in Germany.
- Rogers, Cummins et al. (2023), England: Associations between trajectories of obesity prevalence in English primary school children and the UK soft drinks industry levy: An interrupted time series analysis of surveillance data.
- Cobiac et al. (2024), England: Impact of the UK soft drinks industry levy on health and health inequalities in children and adolescents in England: An interrupted time series analysis and population health modeling study.
- Gračner et al. (2022), Mexico: Changes in weight-related outcomes among adolescents following consumer price increases of taxed sugar-sweetened beverages.
- Basto-Abreu et al. (2019), Mexico: Cost-effectiveness of the sugar-sweetened beverage excise tax in Mexico.

5. Even in cases where the assumptions are fundamentally correct and the input data is representative, the results of modeling studies are always subject to uncertainties that could be considerable, and are often ignored in public health issues.

In order to prevent inaccurate conclusions with potentially far-reaching consequences, these methodological limitations of modeling studies and the resulting uncertainties in their results must be clearly disclosed and taken into account in public and political debates.

Key limitations of the evaluated studies

The key methodological limitations of modeling studies illustrated below are based on the six evaluated studies on the effects of a sugar tax. The analysis is set into three evaluation categories: In the *data foundation* category, the quality of the fundamental, raw data for the modeling studies is evaluated; in the *model specification* category, the models are examined as analytical tools for data processing; in the *communication of results* category the model results are presented and interpreted.

Assessment of the Data Foundation

The samples in the evaluated studies do not always adequately represent the population of interest. Two studies are based on real samples that structurally differ from the overall population of interest: In Rogers, Cummins et al. overweight and obese girls are underrepresented; Gračner et al. in turn, look at exclusively young people who live in urban areas and have health insurance with a particular company. The results of both studies are therefore not reliable. The remaining four studies use synthetic samples derived from current official population statistics (Emmert-Fees et al.; Cobiac et al.; Basto-Abreu et al., Schwendicke and Stolpe), which suggests a basic structural correspondence with the overall population of interest. Moreover, the analysis of Schwendicke and Stolpe from 2017 is based on population data from 2012, which may differ from the current German population due to demographic changes such as migration.

Additional input data in all of the evaluated studies show some considerable qualitative shortcomings, which lead to skewed model parameters and therefore unreliable conclusions. Often, the input data only applies to the application context of the study **to a limited extent**, for example because they originate **from other countries** that do not fit the sample under investigation due to country-specific differences (Emmert-Fees et al.: prevalences of coronary heart disease and strokes from the UK; Schwendicke and Stolpe: price elasticities from the USA; Basto-Abreu et al.: conversion factors from changes in consumption to changes in weight of children from the Netherlands). In addition, some data sets are based on **specific subpopulations** and therefore cannot be generalized to the population as a whole. Nevertheless, Basto-Abreu et al. use generalized data from teachers to convert changes in consumption into weight changes in adults in general. In addition, data is often not available in sufficient **detail** to adequately reflect the heterogeneity of the population. For example, Basto-Abreu et al. assume generalized changes in consumption across all age and gender groups. Furthermore, the data does not always **accurately reflect reality**. For example, data on the purchase of sugar-sweetened beverages in several of the evaluated studies is limited to stationary retail through the exclusion of the gastronomy sector (e.g., in Cobiac et al., Gračner et al., and Basto-Abreu et al.), which leads to distorted results. In addition, some of the data is **outdated**. For example, the consumption data in

Schwendicke and Stolpe and in part also in Emmert-Fees et al. originate from a German national consumption survey (*Nationale Verzehrsstudie II*) from the years 2005 to 2007; the data on medical costs in Emmert-Fees et al. even goes back to 1999 in some cases. In addition, some of the data is based on **small samples**, which limits its reliability. For example, the estimate of productivity losses due to strokes in Emmert-Fees et al. is based on only 151 patients. Another problem was the lower evidence base collected from **non-experimental data**, which many studies rely on due to the lack of experimental data (e.g., conversion factors from consumption to weight changes in Emmert-Fees et al. and Basto-Abreu et al.). If data from **meta-studies** is used (e.g., health data in Cobiac et al.), the quality of the underlying primary data often remains unclear, especially with regard to possible biases and their fit to the study context. Data based on **self-reported data** (e.g., Schwendicke and Stolpe: body indices; Cobiac et al.: Beverage purchases and health data; Basto-Abreu et al.: basic consumption) is prone to errors and can lead to uncertainties in the results. Beyond that, some data sets are **incomplete**, for example Emmert-Fees et al. lack to account for risk factors for coronary heart disease and strokes. Procedures to compensate for missing data (Emmert-Fees et al.: estimation; Basto-Abreu et al.: extrapolation and supplementation with data from other studies) do not always ensure validity. Finally, some datasets are **incompatible** with each other (e.g., consumption data in Emmert-Fees et al.; consumption and body data in Schwendicke and Stolpe), which can lead to deviations in the results, for example in specific age groups. In summary, all studies are based on partially skewed parameters. Modeling on this type of fragile basis – Emmert-Fees et al., for example, use data from 36 different sources to parameterize the model – does not allow any reliable conclusions to be drawn.

Assessment of the Model Specification

The modeling methods used are not suitable for proving causality and therefore do not provide a basis for making the tax *causally* responsible for the simulated effects. The modeling methods used (microsimulation, Monte Carlo simulation, cohort simulation, time series analysis and multivariate regression analysis) are principally suitable for comparing real situations with hypothetical scenarios. However, causality is assumed beforehand in all methods during model construction, but not demonstrated by the model itself. None of the studies explicitly address whether a causal relationship for certain model assumptions has been proven beforehand.

The often inadequate operationalization of the predictor variables leads to a strong generalization of the results. For example, consumption is often simplistically equated with purchasing (Cobiac et al.; Basto-Abreu et al.). In addition, the heterogeneity of consumer behavior with regard to age and gender-specific differences is often not taken into account (Emmert-Fees et al.; Schwendicke and Stolpe; Cobiac et al.; Basto-Abreu et al.).

The simplified assumptions regarding the relationship between influencing and target variables in all evaluated studies suggest that the model results do not accurately reflect real relationships. The relationships are usually modeled linearly and without sufficient empirical evidence. For example, a direct correlation between changes in consumption and weight is assumed (Schwendicke and Stolpe; Cobiac et al.), whereby physiological differences, physical activity and caloric compensation are largely ignored, which contradicts the multifactorial development of overweight and obesity. Studies that model health-related effects (Emmert-Fees et al.; Schwendicke and Stolpe; Cobiac et al.; Basto-Abreu et al.) make assumptions about the effectiveness of the tax over time, but do not take into account fluctuations or long-term adjustment effects in consumer behavior. The other studies also make highly simplifying assumptions: Gračner et al. assume a linear relationship between price changes and body measurements without empirical evidence; Rogers, Cummins et al. assume a linearly continued trend in obesity prevalence. This does not accurately reflect the real correlations.

All evaluated studies take insufficient account of relevant influencing factors, which entails the risk of error prone or highly generalized conclusions about the tax effects. Substitution effects and socioeconomic status are not adequately considered in any study. Similarly, parallel measures such as information campaigns or political reforms are left out. One example is the study by Basto-Abreu et al., which examines the effects of taxing sugar-sweetened beverages in Mexico. Their simulation results are based on the assumption that consumption of these drinks will fall as a result of the tax. However, other studies show that through the inclusion of other food, total calorie purchases in Mexico do not change significantly – which significantly limits the reliability of these results from Basto-Abreu et al.

Assessment of the Communication of Results

The variability of the results is not always transparently presented, and inadequate contextualization may lead to overinterpretation of the simulated effects. Confidence intervals (e.g., as used by Emmert-Fees et al. or Rogers, Cummins et al.) capture only random errors, not systematic errors. Therefore, they only partially reflect the variability of simulated results. In the evaluated studies, adequate contextualization of effects is sometimes lacking. For instance, Emmert-Fees et al. estimate the reduction in healthcare costs across three scenarios to be around three billion euros. This corresponds to less than 0.04 % of total cost – information that is not communicated by the study authors. This omission makes the effect appear disproportionately large. A similar issue arises in studies by Schwendicke and Stolpe, Rogers, Cummins et al., Cobiac et al., and Basto-Abreu et al.

Uncertainties and potential biases of the simulated results are inadequately addressed in all the reviewed studies. Systematic analyses of (global) uncertainties in modeling outcomes are largely absent, and limit themselves mostly to sensitivity analyses. The conducted sensitivity analyses examine variations in individual parameters without altering central model assumptions, such as the underlying influencing factors or assumptions of linear relationships. Given the weak underlying data quality across all studies, these analyses are only partially meaningful and remain questionable.

The model results are often without sufficient consideration of limitations and uncertainties, which makes their reliability questionable and fosters overly generalized conclusions. Although the limitations of the studies are usually discussed extensively, they are often addressed only superficially and rarely reflected upon with regard to their potential impact on the results. For example, Emmert-Fees et al. present the tiered tax as the optimal form of taxation but disregard the uncertainties that were revealed in a sensitivity analysis. Rogers, Cummins et al. and Gračner et al. report effects only for certain subgroups, such as girls, yet generalize these findings to boys. Flawed causal inferences are particularly problematic: Schwendicke and Stolpe, Cobiac et al., and Basto-Abreu et al. conclude that a tax would reduce obesity, even though this is not demonstrated by the model. Such misleading conclusions are especially critical, as they are often taken up by the media and used to justify regulatory measures.

Table of Contents

1 Introduction

Problem Formulation

In discussions about the introduction of a *sugar tax* in Germany (e.g., Appelhans, 2024; Heitmann, 2024; Magoley, 2024; Werny, 2024), reference is made to modeling studies that simulate potential effects of such a tax in Germany or evaluate the effectiveness of the tax in countries that have already introduced the measure. However, the extent to which these studies methodologically meet the necessary quality standards to serve as an evidence basis for well-founded discussions or political decisions has not yet been systematically investigated. The presented expert opinion aims to close this gap. It specifically investigates a selection of studies on possible effects of a tax on sugar-sweetened beverages, some of which have received considerable media presence in the public debate, and evaluates them from a statistical-methodological perspective.

Background

The question „What if...?“ is one of the fundamental considerations in the analysis of potential impacts of measures or decisions. To estimate such effects, modeling studies are used (Jahn et al., 2022). These studies are an established tool to illustrate hypothetical developments and their potential consequences. In so-called *counterfactual* analyses, an observed development is contrasted with a hypothetical one. Comparing the two situations allows, under certain conditions, an assessment of the impact of the implemented measure.

Modeling studies are primarily used to analyze the potential effects of political measures (Münnich et al., 2021), for example, by simulating future societal developments under different scenarios (Schmaus, 2023). In the health sector, they are used to examine the effectiveness of nutritional policies to prevent certain diseases such as diabetes (Mertens et al., 2022). Decision-makers thus use modeling studies as an important source of information and as a basis for evidence-based political decision-making (Jahn et al., 2022) – independent from whether this was originally intended by the researchers or not. However, the statistical-methodological quality of the studies (e.g., the adequacy of the data basis) is not always verified.

If studies are to contribute substantially to political decision-making, they must meet exceptionally high quality standards. They must adhere to the strictest criteria regarding data quality, the suitability of the methodology, and compliance with scientific ethical principles. To date, studies used for political decision-making often do not meet these requirements.

Studies with particularly high quality standards require, among other things, a high degree of accuracy in the results (Baker et al., 2013; European Union, 2020; Münnich, 2023). This means that the parameters estimated using a model should be as close as possible to the actual, usually unknown values (European Union, 2020). Accuracy in the results of modeling studies can only be achieved if both the model assumptions about the real causal relationships are correct and if there is accuracy in the model parameters:

- All modeling studies are based on a (often mathematical) model that represents a complex real situation in a simplified manner (Arnold et al., 2019; Bungartz et al., 2013). Since a complete representation of reality is not possible, the real situation is reduced to the factors considered relevant for the question at hand (Weyer & Roos, 2017). At various points, **assumptions** are made, for example, about which factors are taken into account and how these relate to or interact with each other. Due to these assumptions, the results of modeling studies are always subject to uncertainty (Emmert-Fees et al., 2024).

- Uncertainty in modeling results also arises from the chosen model parameters, which are drawn from various external data sources (Jahn et al., 2022) and incorporated into the model (Emmert-Fees et al., 2024). The **quality of the input data** thus determines the quality of the results (Li & O'Donoghue, 2012). High-quality data is an indispensable prerequisite for evidence-based policy-making (Jahn et al., 2022; Münnich, in press; United Nations, 1954).

Modeling studies rarely explicitly address how underlying assumptions or the quality of the data may influence their results. An example is the recently published modeling study by Emmert-Fees et al. (2024), which thoroughly and specifically examines how strongly the simulated effects of a potential tax on sugar-sweetened beverages depend on the considered factors and the chosen model parameters[6]: The projected changes in body weight and the estimates for health and health economic indicators varied considerably depending on the scenario. This illustrates that (1) the results of modeling studies should be critically examined regarding their dependence on assumptions and (2) these assumptions should be tested for their robustness – especially when they serve as a basis for regulatory measures.

Creating a modeling study suitable for justifying political measures requires high-quality and context-relevant data, realistic assumptions for modeling the relationships between the data, and appropriate statistical procedures. Only on this basis, results that are convincing in their entirety can be achieved. Inaccurate, biased, and/or non-representative data, as well as incorrect assumptions about the underlying causal mechanisms, lead to model results that, while consistent within the model, do not accurately reflect reality.

Contents and Structure of the Expert Opinion

Numerous modeling studies examine the effects of a tax on sugar-sweetened beverages (abbreviation: *SSBs*). The present expert opinion addresses a selection of such studies, which have received considerable media attention in public debate, and evaluates them from a statistical-methodological perspective. For this purpose, the data basis, the specification of the models, and the communication of the results are comprehensively analyzed.

A total of six modeling studies were considered: Central was the evaluation of the study by Emmert-Fees et al. (2023), which examined the effects of different variants of a tax on sugar-sweetened beverages on health and health policy aspects in Germany. This study is frequently cited in the public debate about the potential introduction of such a tax in Germany. Moreover, another study for Germany was evaluated (Schwendicke & Stolpe, 2017). Studies from countries that have already implemented such a tax were also considered. Two of these studies (Rogers, Cummins et al., 2023 and Cobiac et al., 2024) examined the effects of the tax in England, where manufacturers pay varying levies under a tiered consumption tax based on the sugar content of their beverages. Two further studies examined the impact of the tax in Mexico (Gračner et al., 2022 and Basto-Abreu

[6]While the study evaluated in this report by Emmert-Fees et al. (2023) simulates the effects of a potential tax on sugar-sweetened beverages in Germany under fixed assumptions (focus: content), the follow-up study by Emmert-Fees et al. (2024) explicitly examines, in a comparable context, how variations in these assumptions affect the results (focus: methodology). The findings of the more recent study serve as central arguments for the statistical-methodological perspective adopted in this report, particularly in Chapter 2 and Chapter 3. Due to the different publication years of the studies, it is always clearly traceable which study is being referred to in this report.

et al., 2019), where the tax was implemented as a flat consumption tax.

The report is structured as follows: First, a detailed review of modeling studies is presented, including their fundamental potential as well as their limitations (Chapter 2). Based on this, evaluation criteria for modeling studies are derived (Chapter 3), which are subsequently applied to the evaluation of the six selected studies (Chapter 4). Finally, the results of the individual evaluations are summarized and discussed (Chapter 5).

2 Modeling Studies

Modeling studies allow for the contrast between hypothetical and actual scenarios. Modeling studies can be applied to estimate the expected impacts of a potential product tax (Emmert-Fees et al., 2024; Mertens et al., 2022). The results of these models are frequently used by political decision-makers – whether intended by the study authors or not – as a basis for discussions about such measures. An understanding of the general potentials and limitations of modeling studies (Section 2.1) is necessary to assess statements regarding the relationships, causes, and effects of the analyzed topic (Section 2.2).

2.1 Potential and Limitations

The potential and limitations of modeling studies can be summarized as follows:

Potential

- **Contribution to Decision-Making**: Modeling studies can contribute to evidence-based decision-making (Jahn et al., 2022). Ideally, they combine the best available data in a mathematical model to simulate hypothetical scenarios (Emmert-Fees et al., 2024). It should be noted, however, that even the best available data may be of low quality, which limits the contribution of the resulting findings to decision-making (see the detailed discussion in the subsequent *Limitations* section). What "best available" means in a given context should always be carefully examined and communicated transparently. Provided high-quality data, modeling studies serve as a valuable tool for systematically addressing "what-if"-questions and for analyzing the potential impacts of various political measures (Emmert-Fees et al., 2024; Li & O'Donoghue, 2012).

- **Ex-ante Analyses**: Modeling studies analyze the potential impacts of a measure before it is actually implemented (Mertens et al., 2022). In this way, estimates of the expected effect sizes or cost–benefit ratios can be made beforehand.

- **High Level of Detail**: Depending on the chosen methodology, modeling studies analyze not only the potential behavior of the overall population or its subgroups (macro level) but also the potential behavior of individual persons and their interactions (micro level; Schmaus, 2023).

- **Consideration of Complexity**: In principle, arbitrarily complex issues can be modeled, allowing for a comprehensive approximation of reality (Weyer & Roos, 2017).

Limitations

- **Data Availability**: A practical limitation of modeling studies is the insufficient availability of suitable data for initializing and parameterizing the models (Emmert-Fees et al., 2024; Weyer & Roos, 2017). Ideally, models integrate the best available data (Emmert-Fees et al., 2024). However, if suitable (e.g, representative and accurate) data is lacking, this has serious implications for the validity of the modeling results (Mertens et al., 2022).

  In nutritional science applications, sufficient data at the product level or for specific population groups is often lacking, even though there is variability in this regard[7]. Therefore, assumptions must be used as a substitute, which can lead to biases or imprecision in the results.

- **Data Quality**: High-quality data forms a crucial basis for research and evidence-based policy (Jahn et al., 2022). However, poor data quality often poses significant challenges for researchers (Emmert-Fees et al., 2024). Since the quality of the input data directly affects the quality of the results of modeling studies (Li & O'Donoghue, 2012), the validity of modeling results largely depends on the choice of data sources and the uncertainties in the data (Mertens et al., 2022). If data is to be used as a basis for legislative processes, particularly stringent requirements must be imposed on their quality (Münnich, in press).

  Input data collected through surveys of individuals can be subject to various errors. Errors that occur during the design, collection, processing, and analysis of the data may lead to deviations of the estimated values (e.g., means) from the actual values (Biemer, 2010). These include both random errors, such as the *sampling error*[8], and non-random errors such as *coverage error*, *nonresponse error*, *measurement error*, or *processing error* (Lohr, 2021). If, for example, the input data was not collected *probabilistically* (i.e., purely at random) or if nonrespondents systematically differ from respondents, the generalizability of the data is limited (Münnich, 2020). In such cases, the results predicted by the model for the population under study may not hold. Therefore, modeling results based on survey data should be interpreted with appropriate caution (Jahn et al., 2022).

  Modeling studies generally source their input data from external sources and often assume their representativeness and accuracy. However, the absence of a critical discussion of the data foundations suggests a supposed certainty in the validity of the results, even though the risk of bias exists. Data collections on nutritional science topics are particularly susceptible to measurement errors due to misreporting (Emmert-Fees et al., 2024): Individuals tend to not truthfully report their consumption of products that could be perceived as questionable in the context of wanting to provide socially acceptable responses. This can lead to an underestimation of the consumption of such products (Emmert-Fees et al., 2024; Mertens et al., 2022). When such data is used in modeling studies, the results can be significantly biased (Emmert-Fees et al., 2024).

  Data that is considered high-quality in a specific application context does not necessarily have the same quality in another context. For example, representative data on the baseline consumption of Mexican adolescents may be suitable for a study on that target group, but is unsuitable when applied to German adolescents, as physical and sociocultural conditions can vary considerably between the two countries.

- **Assumptions**: Forecasts and projections are based on explicit or implicit assumptions (Ernst et al., 2023), as numerous decisions must be made in the development of models (Schmaus, 2023). Every decision for a specific assumption is simultaneously a decision against alternative possibilities and inevitably introduces uncertainty into the model. Thus, modeling results are significantly influenced by the underlying assumptions (Mertens et al., 2022).

[7] For example, the consumption of sugar-sweetened beverages varies by age, gender, and specific product (Emmert-Fees et al., 2024).
[8] Sampling error arises because only a randomly selected sample is collected instead of the complete population (Faulbaum, 2022; Lohr, 2021).

For example, the method-focused study by Emmert-Fees et al. (2024) showed that the predicted change in body weight associated with the reduction in consumption of sugar-sweetened beverages depends significantly on the assumptions made – for instance, whether caloric substitution by other foods or beverages is considered or whether it is assumed that no such substitution occurs.

- **High Effort**: Modeling studies are often associated with a high level of effort in representing real-world situations in the form of models. As the overall complexity of the model increases, so does the number of potential influencing factors, each of which is subject to uncertainty in its estimation and can therefore increase the overall uncertainty of the modeling results (Schmaus, 2023).

  Furthermore, the potential complexity of the modeling is limited by the computing power of the computer used as well as the available time for implementation and analysis (Weyer & Roos, 2017). The trade-off between effort and benefit is often challenging, as reducing model complexity by simplifying assumptions also introduces uncertainties into the modeling results, as demonstrated by the method-focused study by Emmert-Fees et al. (2024).

  Nutritional science studies require the modeling of highly complex systems, for example, the human body and its responses to environmental influences. It is assumed that numerous nonlinear effects and interactions exist, which cannot be fully captured with limited data and data processing resources.

The above explanations show that the results of statistical models can be subject to uncertainty for various reasons (Jahn et al., 2022; Schmaus, 2023). Quantifying this uncertainty presents a significant challenge (Schmaus, 2023). Competent handling of modeling studies involves the deliberate consideration and transparent communication of uncertainty in the results. Instead of relying on an isolated point estimate, an interval or distribution of possible outcomes should be considered in order to transparently communicate the uncertainties (Jahn et al., 2022).

Modeling studies on public health issues often ignore existing uncertainties in the results, as demonstrated by a method-focused study (Emmert-Fees et al., 2024). This is particularly problematic when the results are used as a basis for discussions about political measures or for their purported fact-based legitimation.

2.2 Effects and Influencing Factors

The modeling studies evaluated in this report examine various impacts (*effects*[9]) of a (potential) sugar tax. In the following, the underlying assumptions about effects in modeling studies (Section 2.2.1) as well as other possible influencing factors (Section 2.2.2) are examined from a methodological perspective.

2.2.1 Relationships, Causes, and Effects

It is important to clearly distinguish between the relationships among variables. **Causality** describes the definitive relationship between cause and effect, i.e. a change in one variable A leads to a change in another variable B. **Correlation**, on the other hand, means that although there is a statistical association between variables A and B, a change in A does not necessarily result in a change in B. Correlation does not automatically imply causality. Furthermore, it is not always clear which variable is responsible for the changes in the other.

[9] An effect is the functional representation of an actual or assumed causal relationship between an influencing factor and a dependent variable.

The recognized gold standard for demonstrating causality are *randomized controlled trials* (Arnold et al., 2019; Jahn et al., 2022). However, due to practical research reasons, they are not suitable for investigating many types of societal issues – including the context of a sugar tax (The Royal Swedish Academy of Sciences, 2021).

When randomized experiments are not feasible and observational data is used instead, drawing causal conclusions becomes considerably more difficult (Jahn et al., 2022). Providing convincing evidence of causal relationships is one of the most challenging tasks in empirical statistics[10].

In modeling studies, the developed models are based on assumptions about underlying (causal) relationships (Jahn et al., 2022). What matters is whether these assumed causalities are based on actually proven or merely hypothetical relationships. Thus, the validity of the results of modeling studies regarding (causal) relationships is limited by how well the causal relationships assumed in the model have been substantiated (Mertens et al., 2022). While modeling studies can be helpful in understanding and explaining (potentially causal) patterns, they do not themselves establish causal relationships in the statistical sense. In plain terms, this means that causal effects between influencing and target variables are incorporated as a prerequisite in the construction of the model – they are assumed beforehand[11]. An assessment of the validity of the causality assumption is made retrospectively and indirectly, when the model's simulated results sufficiently match reality. However, this does not constitute (scientific) proof, but rather a plausibility of the assumptions.

To "prove" causal relationships in the field of public health, mainly counterfactual analyses are employed (Arnold et al., 2019). These are, for example, simulation studies that compare the impacts of public health measures – such as introducing a tax on sugary foods with the intention of reducing sugar consumption in the population and the incidence of related diseases – with the scenario without such a measure. Even during the construction of the model, assumptions about underlying (causal) relationships are made. For instance, it is assumed that price changes in taxable products are the cause of changes in the consumption of these products, or that a reduction in the consumption of taxed sugary foods leads to weight loss. Even if the results based on these assumptions appear plausible, without corresponding evidence, a definitive causal relationship cannot be established.

2.2.2 Unaccounted Influencing Factors and Isolated Effects

Verifying and interpreting (causal) relationships in data is also challenging because observed associations may be influenced by omitted variables that are not considered in the model (Jahn et al., 2022).[12]

To accurately assess the true impact of planned interventions, such as a sugar tax, its isolated effect must be ascertained as precisely as possible. In statistics, this problem is addressed by including additional potential

[10] The complexity of this problem was also highlighted by the awarding of the Alfred Nobel Memorial Prize in Economic Sciences in 2021, half of which was awarded to researchers recognized for their methodological contributions to the analysis of causal relationships. They examined the question of what conclusions about causality can be drawn from so-called *natural experiments* (The Royal Swedish Academy of Sciences, 2021).

[11] In modeling studies examining the effects of a sugar tax, the following (causal) relationships are typically assumed beforehand: The starting point is the (hypothetical) introduction of such a tax, which – depending on its design – can trigger different market reactions. Manufacturers respond with price increases or recipe adjustments, leading to changes in the purchasing and consumption behavior of the population. A reduced consumption of sugar-sweetened beverages can lower overall sugar intake, which in turn can lead to reduced caloric intake and potential changes in body weight. These changes, in turn, affect health, quality of life, and health-economic aspects.

[12] The literature distinguishes (sometimes inconsistently) various such third-variable effects (e.g., spurious correlation, interaction, confounding, mediation, moderation; Weiber and Mühlhaus, 2014). These effects differ in whether and how the third variable is related to the dependent and/or independent variable. A more detailed differentiation of third-variable effects is not necessary for the purposes of this report.

influencing factors as control variables alongside the presumed causal factor (Arnold et al., 2019). This allows the influence of the control variables on the dependent variable to be separated out (Doering & Bortz, 2016). However, a major challenge remains in adequately accounting for all relevant influencing factors. In the case of a sugar tax, potential substitution effects (e.g., consumption of other high-calorie beverages or foods) would need to be incorporated into the model. Similarly, all other risk factors that promote the development of obesity must be modeled at the individual level. Obesity is considered a multifactorial disease, influenced not only by dietary habits but also by numerous other factors such as the social environment or the level of physical activity (Hummel et al., 2013; World Health Organization, 2022). If these factors are not taken into account, there is a risk that modeling studies will lead to serious inaccurate conclusions regarding the potential impacts of such a tax.

3 Evaluation Criteria for Modeling Study Assessments

Based on the characteristics of modeling studies described in Chapter 2, this chapter derives the evaluation criteria for the studies, which are divided into three categories: *Data Foundation*, *Model Specification*, and *Communication of Results*. Data serves as the raw material, the models are the analytical tools for processing it, and the results represent the resulting product.

3.1 Data Foundation

The data foundation is often insufficiently discussed or not discussed at all in modeling studies. However, since high-quality data is an indispensable prerequisite for evidence-based policymaking, its evaluation plays a central role in this report.

Availability and Quality of the Underlying Input Data

As explained in Section 2.1, the accuracy of the results of modeling studies depends significantly on the availability and quality of the underlying input data, which typically come from various external sources. To obtain reliable study results, the relevance of the data to the research subject must be ensured, and the risk of bias minimized. The input data of the evaluated studies are assessed with regard to their up-to-dateness, completeness, data collection methods (e.g., surveys, observations, experiments), the number of data collection points (e.g., cross-sectional or longitudinal[13]), potential random and non-random errors (see Section 2.1), and their applicability in the study context[14]. Additionally, the compatibility of the various data sources is evaluated.

For example, when examining the potential impacts of taxing specific products, it is crucial to determine how the *price elasticity of demand* evolves, i.e., the demand for a good in response to price changes. This is generally unknown and must therefore be determined in studies. However, the results are typically subject to uncertainty or are not representative. For instance, a price elasticity measured exclusively in urban areas cannot automatically be applied to rural regions. Furthermore, it must be examined whether the underlying data is suitable for the spatial, temporal, and contextual framework of the respective study. Empirically derived price elasticities cannot be directly transferred to other countries, as they depend on numerous other factors, such as the price level of substitutes. The choice of survey methodology also plays a crucial role. Consumption estimates based on individual surveys differ from estimates based on sales figures (Emmert-Fees et al., 2024). Sales figures themselves vary depending on the considered distribution channels, such as whether consumption is measured solely based on retail data (e.g., supermarkets) or including data from the hospitality sector (e.g., restaurants, cinemas).

Samples

Within the assessment of the data foundation of a study, it is especially important to consider how well the underlying sample represents the population of interest. Samples in modeling studies can either be *real* or

[13] In a *cross-sectional study*, a sample is examined at a single point in time. A *trend study* consists of several temporally staggered cross-sectional studies, where at least some of the same variables are recorded. These studies are used to analyze societal changes over time. A *longitudinal or panel study* repeatedly examines the same sample (panel) over an extended period. It is particularly suitable for analyzing individual changes over the course of life (Doering & Bortz, 2016).

[14] Especially when the data were not self-collected, examining potential biases can be a significant challenge. Indications are provided, in particular, by descriptions of the data collection process and handling nonresponse data.

synthetic:

- Real samples are subject to certain fluctuations, even with random selection and large sample sizes, which are referred to as (random) sampling errors[15]. Confidence intervals are typically provided to represent the extent of random errors. However, these can only be correctly determined if probabilistic[16] samples are present.
- Synthetic samples are artificially created datasets that closely replicate a population using parameters of the population distribution and demographic characteristics in order to simulate specific scenarios. The results of a model based on such samples depend on how well the sample reflects the characteristics of the population being considered and the assumptions made about the course of life of the individuals being studied (Mertens et al., 2022).

The composition of a sample must allow conclusions to be drawn about the entirety of the population of interest at the relevant time. This requirement is referred to as the *representativeness*[17] of a sample and is a decisive factor whether and how the study results can be generalized. The samples of the studies evaluated in the report are therefore assessed in terms of their representativeness.

3.2 Model Specification

To assess the model risk – the risk arising from a model's deviation from reality – the expert opinion examines the specifications of the respective models used. The establishment of certain model assumptions (and thereby the simultaneous rejection of alternative assumptions) lead to uncertainty in the model, which can be traced back to various aspects of model specification:

Model Methodology

The results of the modeling studies are determined by the modeling methods and statistical analyses, the appropriateness of which depends on the respective research goals. To assess the validity of the results, the limitations of the respective methods (see Appendix A) must be taken into account.

Influencing and Target Variables

The structure of the model is defined by specifying the influencing variables (independent variables) and target variables (dependent variables). This includes, on the one hand, the operationalization of these variables, i.e., how theoretical constructs (e.g., quality of life) are translated into measurable variables (e.g., years of life, years lived with disease; Doering and Bortz, 2016). On the other hand, it is important how these variables are related. These relationships are described by functional connections between the influencing and target variables, which

[15] In general, the larger the sample, the smaller the sampling error for unbiased samples. In practice, this means that larger, unbiased samples provide more accurate estimates (Mittag & Schüller, 2023). It is particularly important to critically assess relatively small sample sizes that deal with subpopulations, such as certain age groups (Li & O'Donoghue, 2012).

[16] A probabilistic sample requires that the probability of each individual being included in the sample is positive and known. Only such samples can claim representativeness, with adjustments to the overall population possibly made beforehand (by quota sampling) and/or afterwards (through weighting).

[17] The interpretation of the term *representativeness* is by no means trivial. In particular, the representativeness of a sample alone is not enough to ensure high-quality results. To ensure this, both sampling and non-sampling errors must also be considered (Münnich, 2020).

can be modeled either linearly or nonlinearly[18]. Interactions between influencing factors may also be relevant. If the model structure does not adequately represent the real relationships between the dependent variables, there is a risk of error-prone conclusions.

Other Influencing Factors

To minimize biases in the results, a model should not only include the central influencing and target variables, but also other variables that are related. If such relevant influencing factors are not sufficiently considered in the model, there is a risk that relationships between influencing and target variables will be derived from results that are overly generalized or incorrect. The influencing factors that are actually relevant in the context of the sugar tax cannot be definitively determined at present. However, some significant factors affecting consumption behavior and consumption changes due to price changes are identified in the methods-focused study by Emmert-Fees et al. (2024):

- The base consumption of sugar-sweetened beverages, i.e., consumption before a (potential) taxation of these products, varies by age and gender. Additionally, individuals with different base consumption levels respond differently to the introduction of a tax, which is reflected in varying price elasticities of demand in different age and gender groups. Therefore, assuming a uniform base consumption and a uniform price elasticity of demand when modeling the effects of a tax introduction would be an oversimplification. Instead, at least *age* and *gender* should be included as control variables[19].
- The change in price of a good (e.g., through taxation) can affect the consumption of other goods, which is referred to as *cross-price elasticity*. For example, sugar-sweetened beverages may be partially substituted by untaxed caloric drinks (e.g., fruit juices, milk-based drinks, or also alcoholic beverages) after a price increase[20]. Therefore, when modeling the effects of a tax introduction, it is important to control for the potential caloric compensation (Hummel et al., 2013) – an aspect often overlooked (Thiboonboon et al., 2024) – otherwise, there is a risk of significantly overestimating the actual effect.

3.3 Communication of Results

The presentation and communication of results is a central component of any study, including modeling studies. In this context, not only is the content precision important, but also the way in which the results are conveyed, as studies may be received by different audiences, such as scientists, the media, policymakers, or the general public.

Presentation of Results

The presentation of results conveys – especially in quantitative studies – the central finding of a scientific investigation. To ensure objectivity, reproducibility, and reliability, the results of a study must be presented in a neutral, complete, transparent, consistent, understandable, and replicable manner. If these principles are not

[18] Examples of nonlinear functional relationships include sigmoidal or exponential relationships.

[19] Modeling the effect of price changes may fall under model risk if price elasticities are not differentiated by subgroups in the model. On the other hand, it can also be part of the sampling risk if the studies on which the estimated price elasticities are based contain from high random or systematic errors.

[20] As like the price elasticities of demand, the cross-price elasticities of certain substitutes for sugar-sweetened beverages are related to the base consumption of these drinks (Emmert-Fees et al., 2024).

adhered to, there is a risk that the results will be misunderstood, misinterpreted, or incorrectly communicated.

An incomplete presentation, such as selectively presenting significant results from certain subgroups, can obscure non-existent or unverifiable effects. The visualization of results also requires special care, as the choice of presentation form (e.g., pie charts or bar graphs) and the underlying scales can emphasize certain statements, thus steering the reader's perception in a particular direction. The way study results are presented can consciously or unconsciously influence how they are received. Likewise, the choice of presenting results in absolute or relative numbers can convey different messages. Relative frequencies, when based on small initial values, tend to make risks appear larger, so that their mere indication can be potentially misleading. For this reason, Wegwarth and Gigerenzer (2011) and Krauss et al. (2020) advocate for communicating risks and risk changes in natural frequencies ("x out of y"). It is similarly difficult or misleading when projections depict absolute numbers without any contextualization, especially when uncertainties are not communicated. For instance, projected sums of national savings for a country with 80 million inhabitants over a period of twenty years are at best rough estimates that require contextualization. In general, a consistent presentation in both absolute and relative values can contribute to the improvement of the interpretability of the effects.

Model Evaluation

The evaluation of the quality and reliability of simulation models is based (especially in microsimulation studies) on three types of analyses (Mertens et al., 2022):

- Validity Analyses: Validation of the model is necessary to substantiate its credibility in terms of its similarity to reality (Li & O'Donoghue, 2012). The validity analysis ensures the alignment of the modeling results with other observed data or theoretical considerations (Mertens et al., 2022). It examines whether the model correctly represents the real phenomenon, for example, through appropriate assumptions, and provides insights into the reliability of the results. Sources of error can occur at various points in the process, such as in the model specification, algorithms, code, or even in the interpretation of results (Bungartz et al., 2013).

- Sensitivity Analyses: Provides information on the robustness of a model, which is relevant for decision-making based on the modeling results. Sensitivity analyses determine the variability of target variables as a function of the variability of influencing factors and thus show how much the results are affected based on the uncertainty in the input values (Saltelli et al., 2004). In practice, it is typically analyzed how the variation of model parameters, which are difficult or not plausible to estimate, affects the results (Mertens et al., 2022).

  Sensitivity analyses thus serve to both identify influential parameters and to estimate their impact on the results. This allows for the comparison of the effects of different factors on a selected target variable (Schmaus, 2023), identifying the parameters that have the greatest impact on the outcome.

  For example, the base consumption of sugar-sweetened beverages is operationalized differently in studies. Some use sales reported by the industry, while others use the consumption reported by consumers[21]. Sensitivity analyses that vary such assumptions can reveal uncertainties in the simulation results arising

[21] Emmert-Fees et al. (2024) report in their methods-focused study that the sales of sugar-sweetened beverages reported by the industry were approximately 1.86 times higher than the consumption reported by consumers themselves.

from the specific operationalization of the parameters (see e.g., Emmert-Fees et al., 2024).

- Uncertainty Analyses: They quantify the (global) uncertainty in the model (Saltelli et al., 2008), which can arise from inaccurate or incomplete input data as well as underlying model assumptions.

  The degree of uncertainty in the assumptions is decisive for the level of evidence that a modeling study can provide. Therefore, the quantification of uncertainty in the model is an essential foundation for decision-making (Mertens et al., 2022).

These three described analyses complement each other to allow for a comprehensive evaluation of the quality and reliability of a model. Model evaluation provides transparency regarding how much of a reliable basis the model offers for decision-making.

Contextualization of Study Results

The transparent contextualization of the results of the study, including the strengths and weaknesses, is essential to draw informed implications for theory and practice. In this process, the limitations of a study (e.g., a lack in the data quality, uncertainties in model parameters, unconsidered influencing factors, or a non-representative sample) must be clearly communicated. This enables a realistic assessment of the significance of the results and their correct interpretation. Especially in a public-facing publication, inadequate communication of limitations carries the risk of a shortened, misleading, or incorrect reception of the results. The study authors bear the special responsibility of providing a well-founded basis for critical discussions, especially when the results are expected to be far reaching or receive broad public perception. This includes, for example, the precise differentiation between correlational and causal relationships, as well as an assessment of the generalizability of the findings. At the same time, potential recipients must be aware that reproduction of results without taking into account their respective limitations can lead to misinterpretations.

Summary of the Assessment Criteria for the Evaluation of Modeling Studies

Data Foundation

- Examination of the **data sample** regarding its size and composition to assess the extent to which it structurally aligns with the whole population of interest in key demographic parameters (a prerequisite for representativeness), and whether the sample as a whole and within the relevant subgroups is sufficient to keep the random error of the estimation of the effects within acceptable limits.
- Examination of **other input data** of the model in terms of their timeliness, completeness, data collection methods and time points, possible random and systematic errors, their applicability in the study context, and their compatibility with each other, to assess how well they fit to the research topic.

Model Specification

- Examination of the **model methodology** and statistical analyses in terms of their suitability and respective limitations, to determine how appropriate they are for achieving the research goal.
- Examination of the **influencing and target variables** in consideration to their operationalization and functional relationships, to assess whether the real phenomena being studied are sufficiently accurately represented.
- Examination of **other influencing factors** in terms of their consideration and relevance in the specific application context, to assess whether there is a risk of errors in conclusions, such as from the failure to account for confounding factors regarding the modeled relationships and effects.

Communication of Results

- Examination of the **presentation of results** in terms of anomalies regarding neutrality, completeness, transparency, consistency, understandability, and replicability, to assess the extent to which the results are presented objectively and with minimal risk of misinterpretation.
- Examination of the completeness of the **model evaluation** in terms of the discussion of validity, sensitivity, and uncertainty, to assess how well the quality and reliability of the model are checked.
- Examination of the **contextualization of study results** in terms of their generalizability and disclosure of study limitations, to assess how well a solid basis for critical discussions is established and how well implications for theory and practice have been addressed.

4 Evaluation of Modeling Studies on the Effects of a Sugar Tax

4.1 Study by Emmert-Fees et al. (2023), Germany

Full reference source:

Emmert-Fees, K. M. F., Amies-Cull, B., Wawro, N., Linseisen, J., Staudigel, M., Peters, A., Cobiac, L. J., O'Flaherty, M., Scarborough, P., Kypridemos, C., & Laxy, M. (2023). Projected health and economic impacts of sugar-sweetened beverage taxation in Germany: A cross-validation modelling study. *PLOS Medicine*, *20*(11), Article e1004311. https://doi.org/10.1371/journal.pmed.1004311

Executive Summary

The study by Emmert-Fees et al. (2023) investigates the potential impacts of a sugar tax on sugar-sweetened beverages in Germany through the use of a microsimulation approach. Three scenarios are modeled to estimate the occurrence of various diseases over a 20-year period. In all scenarios, the simulation results in a reduction in disease incidence as well as in the estimated associated costs, through the introduction of such a tax. However, the proportion of potentially reducible disease cases and healthcare costs relative to total diseases or costs is very small. The savings in the healthcare costs amount to less than 0.04 % in all scenarios. The methodological approach of the study is generally understandable, but there are criticisms regarding neglecting to consider relevant factors. These include substitution effects, important influencing factors (beyond age, gender, and BMI), and heterogeneity in price elasticities. Additionally, the underlying dataset has significant weaknesses that negatively impact the reliability of the model results. These deficiencies include outdated and incomplete datasets, small sample sizes, and the lack of representativeness in the data. Further criticism concerns the presentation of the model results, as it suggests stronger effects and greater certainty than what can actually be derived from the model. Overall, from a statistical-methodological perspective, the study does not meet the necessary quality standards to serve as a reliable basis for evidence-based decisions.

4.1.1 Summary of the Study

The contents of the study by Emmert-Fees et al. (2023) are summarized below.

Research Focus

The study simulates the potential effects of a hypothetical tax on sugar-sweetened beverages[22] in Germany based on three scenarios:

(I) Introduction of a 20 % ad valorem excise tax[23] on sugar-sweetened beverages,

(II) Introduction of a 20 % ad valorem excise tax on sugar-sweetened beverages and fruit juices[24],

[22] Soft drinks and fruit juices with added sugar (Emmert-Fees et al., 2023, S. 6).

[23] The tax is added to the sale price, and the final sale price is increased accordingly.

[24] This is intended for fruit juices, fruit nectars, and other juices (e.g., vegetable juices) that naturally contain sugar. These juices

(III) Introduction of a tiered tax[25], which is assumed to lead to an average reduction of 30 % in the sugar content of sugar-sweetened beverages.

The focus of the simulation is on the health impacts in terms of the development of the incidence of various diseases (Type 2 diabetes mellitus (short: *T2DM*), coronary heart disease (short: *CHD*), strokes, and obesity) in the German population aged 30 to 90 years for the period 2023 to 2043. Additional focus is placed on predicting the associated health-economic impacts in terms of national economic costs.

Methodology

The study uses a microsimulation approach to simulate relevant parameters for 200,000 synthetic individuals (see Chapter 3). This population replicates characteristic features of the German population aged 30 to 90 years and is modeled for the period from 2013 to 2043[26] using census data and population forecasts from the Federal Statistical Office. The results of the simulation are then extrapolated to the total population of this age group for the respective simulation years.

In scenarios (I) and (II), the effect of the respective tax on the consumption of sugar-sweetened beverages and fruit juices is modeled under the assumption of unchanged sugar content of these beverages by calculating price elasticities. In scenario (III), the effect of the tax is modeled under the assumption of a 30 % reduction in sugar content due to recipe changes with unchanged consumption of sugar-sweetened beverages.

In all scenarios, the general consumption of sugar-sweetened beverages as well as the specific consumption of sugar from sugar-sweetened beverages and fruit juices (in addition to other parameters such as age, gender, and BMI) are assumed as individual risk factors for modeling the different disease incidences. These are predicted using a Monte Carlo simulation. The existence of preexisting medical conditions is partially taken into account – see Emmert-Fees et al. (2023), Figure 1, for a systematic representation of the dependencies. From this, prevalences and mortalities are calculated by modeling individual risks, which ultimately form the basis for estimating the effects on quality of life (measured in QALYs[27]) and the societal costs caused by the diseases under consideration.

Results

The simulation shows a reduction in the average sugar consumption from sugar-sweetened beverages and fruit juices in all three scenarios. The greatest average reduction in sugar consumption is simulated at 5.9 grams per day, 95 % CI[28] [5.4; 6.0] in Scenario (II), followed by 2.3 g/day, 95 % CI [2.3; 2.4] in Scenario (III), and 1.0 g/day, 95 % CI [0.1; 1.7] in Scenario (I). The remaining model results provide an inconsistent picture regarding the effectiveness of the individual scenarios: Scenario (II) shows the largest effects in terms of the reduction or delay of strokes and obesity, while Scenario (III) shows larger effects in the prevention or delay of T2DM and coronary heart disease. Scenario (II) has a greater impact on quality of life, while Scenario (III) has larger effects on life

can also contain added sugar (Emmert-Fees et al., 2023).

[25]The fee increases in tiers with respect to the sugar content of a drink, and is intended to motivate manufacturers to adjust the recipes of their sugar-sweetened beverages.

[26]The simulation period consists of a calibration period (2013-2023), during which various model parameters are estimated based on observed data, and a forecast period (2023-2043), during which the effects of the scenarios are predicted.

[27]QALY is the abbreviation for *quality-adjusted life year*. It is a metric used to measure the benefits of medical treatments or health interventions, taking into consideration the lifespan and quality of life of the affected individuals.

[28]A confidence interval (CI) is a random interval that, given a known sample distribution, has a predefined coverage probability for the true but unknown parameter. This means that with a given confidence level of 95 %, which is typically chosen as standard, 95 % of the confidence intervals derived from samples would cover the true value.

expectancy as well as on health and societal costs.[29]

Conclusions of the Study Authors

The study authors interpret the model results in such a way that all examined tax scenarios could result in substantial health benefits as well as significant cost savings. A levy on sugar-sweetened beverages depending on the sugar content presumably achieves the strongest effects in comparison to the other scenarios. The inclusion of fruit juices in a taxation may be associated with further health benefits.

4.1.2 Evaluation of the Study

Data Foundation

The selection of the synthetic sample consisting of 200,000 individuals aged 30 to 90 years based on current official population data is methodologically appropriate. Although a simulation of the entire population could eliminate the potential error source from sampling and increase the accuracy of the estimates, the chosen sample size appears sufficient to generate acceptable estimates. The exclusion of children, adolescents, and young adults from the simulation is justified by the study authors due to the insufficient availability of relevant data for these age groups.

A reliable simulation of the potential impacts of a hypothetical sugar tax in Germany is however complicated due to significant limitations in the availability and quality of the necessary input data.

For the modeling of the distributions of consumption behavior, BMI, and cardiovascular disease incidence, data from the KORA study (Helmholtz Zentrum München, 2024) from the S4 (1999), F4 (2007), and FF4 (2014) cohorts are primarily used, as more recent (and thus more representative) data was not available for Germany according to the study authors.

Data on consumption behavior is only available from the most recent KORA cohort, which does not include individuals under 38 years of age. Therefore, Emmert-Fees et al. (2023) use additional data from a German national consumption survey called *Nationale Verzehrsstudie II* (NVS II; Max Rubner-Institut, n. d.) to model consumption behavior. However, the consumption data from the two sources strongly differ due to different data collection methods (see also Emmert-Fees et al., 2023, Appendix S1, Figures D und H). Emmert-Fees et al. (2023) estimate consumption based on both datasets, but it remains unclear which of the two data sources more realistically depicts consumption.

Other model parameters are based on studies with very small sample sizes[30], which considerably limits the reliability of the data. Some of the data sources are already outdated[31] or lack the necessary variables, which are then estimated based on other variables[32]. This adds further uncertainty to the model results. For other

[29] The simulated differences can likely be attributed to the fact that a reduction in sugar consumption from sugar-sweetened beverages tends to affect different population groups than a reduction in sugar consumption from decreased fruit juice consumption, as the consumption of sugar-sweetened beverages and fruit juices shows significant age and gender specific variations (see Figures D-J in Appendix S1 in Emmert-Fees et al., 2023).

[30] For example, to estimate productivity losses due to strokes, results are based on 151 patients from Winter et al. (2008).

[31] Consumption data based on the NVS II is from the years 2005-2007 (Max Rubner-Institut, n. d.), data on disease costs from Winter et al. (2008) and Icks et al. (2013) are from 1999.

[32] For example, the risk factors for CHD and strokes are not recorded in the used dataset and are instead estimated from those for cardiovascular diseases.

model parameters, no suitable data was available for Germany[33], so supplementary data sources were used, particularly from the USA and the UK. However, the transferability to the German population is questionable, as country-specific differences in manufacturer and consumer behavior, healthcare, and other relevant factors must be considered.

The model parameters, such as for estimating the effects of consumption and BMI on disease incidences, mainly come from observational studies (mostly cross-sectional studies) that provide a lower evidence base when compared to intervention studies. In particular, cross-sectional studies are only able to show correlations and not causal relationships. This limitation results from the lack of availability of intervention studies within this application context. This methodological issue in the study by Emmert-Fees et al. (2023) was also addressed by Nawroth and Kumar (2024) and discussed in a reply by the study authors (Laxy & Emmert-Fees, 2025).

For the parameterization of the model, the data comes from a total of 36 different sources. A complete overview of these data sources is provided in Emmert-Fees et al. (2023), Appendix S1, Table A. This large amount of different data sources, which sometimes use different definitions and data collection methods for the individual parameters, by itself, introduces a large degree of uncertainty into the model.

Model Specification

Microsimulation models are generally the method of choice for simulating complex relationships in large cohorts (see Appendix A.1.2). However, the validity of the model results heavily depends on the availability of reliable data for the individual sub-coherencies and how much the model results are influenced by the underlying assumptions and input parameters. It should also be emphasized that a microsimulation model cannot prove causal relationships. Causality is assumed in the model and must therefore be adequately established for each simulation step in advance. The present study, for methodological reasons, cannot prove a causal effect but can only provide an estimate of possible effect sizes, if such a causal relationship exists.

Regarding the model specification, the following criticisms should be noted:

- Neglecting the heterogeneity of price elasticities: Individual price elasticities can be influenced by various factors, such as the average personal consumption of sugar-sweetened beverages (Blake et al., 2019). However, in this study, it is assumed that with price increases, all consumers would reduce their consumption in the same way. This leads to a significant generalization of the study results.

- Generalization of sugar reduction in sugar-sweetened beverages: In scenario (III), a fixed reduction of 30 % in the sugar content of all sugar-sweetened beverages is assumed as a result of recipe changes. However, deviations from this are expected in reality, which may influence the simulated health and health-related effects, as the study authors themselves explain in a sensitivity analysis[34].

- Failure to consider temporal trends: Temporal developments regarding price elasticities, consumption behavior, the sugar content of sugar-sweetened beverages (except for the formulation change in scenario (III)), and healthcare are not considered. Therefore, the model results likely do not adequately reflect real

[33] For example, data on tax pass-through rates, the prevalence of CHD and strokes and correlations of consumption behavior with BMI, CHD, and T2DM for Germany, is missing.

[34] Detailed explanations of the sensitivity analysis are provided in the section *Communication of Results*.

phenomena and their relationships.

- Limited consideration of influencing factors: Aside from age, gender, and BMI, no other potential influencing factors are considered. However, there could be other important determinants such as physical activity and socioeconomic status (e.g., education level, household income). In particular, substitution effects are only considered in relation to fruit juices. Thus, the effects simulated by Emmert-Fees et al. (2023) do not necessarily exist in the presented form.

Communication of Results

The model results are primarily presented as absolute totals over the 20-year simulation period, without relating them to the respective baseline. This carries the risk of misperceptions and, consequently, unrealistic expectations regarding the potential effects of a tax. An example is the estimated reduction in healthcare costs, which ranges from 2.26 billion euros, 95 % CI [1.19; 3.60], in Scenario (I) to 3.85 billion euros, 95 % CI [2.07; 6.08], in Scenario (III).

Assuming a uniform distribution of savings over the 20-year simulation period, corresponds to annual savings of approximately 110 million euros in Scenario (I) to 190 million euros in Scenario (III). For comparison, the total costs of the German healthcare system in 2022 were 498 billion euros, according to the Federal Statistical Office (2024). The cost savings calculated by Emmert-Fees et al. (2023) resulting from taxing sugar-sweetened beverages would correspond to a relative cost reduction of about 0.023 % in Scenario (I) and 0.039 % in Scenario (III) – a share that can be considered low in the context of total costs.

The conclusions of the authors are mostly cautiously worded (e.g., with phrases like “could lead to”, “likely to lead to”), but some phrases (e.g., “would lead to”) suggest certainty that is not in line with the limitations of the model.

Overall, the model results and the conclusions drawn from them are presented in the study as considerably more reliable than is justified. For example, the authors state: “We comprehensively consider implications of all sources of uncertainty from parameter uncertainty to the included risk relationships and the chosen simulation method [...]. This makes our findings particularly robust“ (Emmert-Fees et al., 2023, p. 17). While certain uncertainties in the input data are appropriately modeled, the consideration of other potential uncertainties remains incomplete or is entirely omitted. For example, uncertainties arising from assumptions about the methodology (e.g., for calculating price elasticities) are not considered. Furthermore, the uncertainty regarding the impact of a tiered tax on the sugar content of sugar-sweetened beverages in Scenario (III) is not modeled, and the non-representativeness of various datasets as well as uncertainties in the societal costs due to early death are not considered. Additionally, only very small ranges of variation are assumed for disease costs.

This limits the validity and generalizability of the obtained model results. Since these sources of uncertainty were not considered, the confidence intervals presented in the study, which only reflect uncertainty from random but not from systematic errors, do not appropriately reflect the actual variability.

Various sensitivity analyses (see Appendix A.3.2) are conducted in the study on the underlying model parameters of the tax scenarios[35] (see Emmert-Fees et al., 2023, Appendix S1, Table V). One of the sensitivity analyses (Sensitivity Analysis 4) shows, for example, that the model results for a tiered tax exhibit significant sensitivity to the assumed reduction in the sugar content of sugar-sweetened beverages: The analysis is designed similarly to Scenario (III), but assumes only a 10 % reduction in sugar content in the beverages. The simulated effects are consistently lower than in the three main scenarios. A detailed comparison of the sensitivity analysis results with Scenario (III) suggests that the reduction in sugar content is nearly linearly correlated with the model results. This implies that even a deviation of one-fifth from the assumed sugar reduction in the beverages (a reduction of 24 % instead of 30 %) leads to relative uncertainties in about one-fifth of the model results. This makes the evaluation of Scenario (III) as a recommended strategy questionable, as the resulting effects would be comparable (in terms of T2DM and coronary heart disease incidences and derived values) or even worse (in terms of obesity and stroke incidences and derived values) than those of Scenario (II).

Sensitivity analyses serve to estimate uncertainties. A more prominent presentation of the analyses conducted in the main text of the study, as well as an expansion of the sensitivity analyses to other sources of uncertainty in the modeling process, would therefore have been appropriate.

A follow-up study from a similar group of authors addresses the uncertainties in the simulation of Emmert-Fees et al. (2023): Emmert-Fees et al. (2024) analyze in a comparable context how different assumptions (e.g., (non)consideration of substitution; specific design of the tax) and model parameters (e.g., price elasticities of demand) influence the estimated effects of taxing sugar-sweetened beverages. The pronounced sensitivity of the simulation results to certain assumptions and model parameters raises significant doubts about the robustness of the model results presented in Emmert-Fees et al. (2023) and the conclusions drawn from them.

Emmert-Fees et al. (2023) nonetheless transparently point out several limitations of their study[36], but some of the limitations are either inadequately mentioned or entirely overlooked. This includes, in particular, the (possible) non-representativeness of studies on (1) tax pass-through rate, (2) sugar reduction in sugar-sweetened beverages through reformulation of recipes, and (3) the impact of sugar consumption on BMI, as well as the neglect of energy substitution when the consumption of sugar-sweetened beverages/fruit juices decreases and is replaced by other caloric liquid and solid foods.

[35] Height of the ad valorem consumption tax, (no) substitution by fruit juices, lesser degree of formulation change by manufacturers, combination of tiered taxes and ad valorem consumption tax.

[36] Mentioned limitations: Non-representativeness of the base data from the KORA study; no modeling of temporal trends in consumption behavior; no consideration of effects beyond diabetes and cardiovascular diseases; no modeling of long-term sugar consumption; no consideration of the heterogeneity of price elasticities; no consideration of the inflation crisis of 2022/23; no modeling of people under 30 years old; no precise modeling of a tiered tax on the sugar content of the affected sugar-sweetened beverages; no consideration of the costs of implementing the taxes and reformulation by manufacturers; non-representative or outdated data for incidences and prevalences; assumptions about sugar content; non-consideration of other relevant influencing factors; very small sample sizes.

Summary of the Limitations of the Study from Emmert-Fees et al. (2023)

Data Foundation

- **Data Sample:** Replication of the German population (limited to 30 to 90-year-olds) based on current official population data is generally assumed to be representative, but is not applicable to children and young adults.
- **Other Input Data:** Data from a total of 36 data sources; partly outdated (consumption data from 2005-2007, disease expenditure data from 1999); partly based on very small samples (e.g., productivity losses from strokes based on 151 patients); transferability to Germany is not always guaranteed (e.g., stroke prevalence); limited compatibility of datasets (consumption data); missing variables (e.g., risk factors for strokes); mainly based on observational studies (e.g., for estimating the effect of consumption and BMI on disease incidences); therefore, overall limited validity of the model results.

Model Specification

- **Model Methodology:** Causality is assumed beforehand in microsimulation models (as here based on Monte Carlo simulations) during model construction, but not demonstrated by the model itself.
- **Influencing and Target Variables:** Failure to consider the heterogeneity of price elasticities leads to significant generalization of the results; due to numerous other assumptions (e.g., 30 % reduction in sugar content for all beverages in Scenario (III), uniform consumption behavior, uniform distribution of disease incidences and costs), the model results likely do not adequately reflect real phenomena and their relationships.
- **Other Influencing Factors:** Inadequate consideration of substitution and failure to account for other relevant factors (e.g., physical activity, socioeconomic status), meaning that the modeled effects do not necessarily exist in the current presented form.

Communication of Results

- **Presentation of Results:** Despite transparent reporting, the confidence intervals only have limited significance; the exclusive presentation using absolute sums without appropriate contextualization creates the impression of stronger effects than actually present.
- **Model Evaluation:** Sensitivity analyses put into perspective the results of the main analysis, and question them; only few sensitivity analyses and lack of examination of the impact of key model assumptions make it difficult to assess model-based uncertainty.
- **Contextualization of Study Results:** Inadequate discussion of limitations; oversimplified classification of model results (e.g., presentation of the manufacturer fee as the optimal solution without considering variability); robustness of the model results is questionable due to uncertainties; conclusions suggest higher certainty than can be derived from the model results.

Overall Assessment of the Study by Emmert-Fees et al. (2023)

The methodological approach of the study by Emmert-Fees et al. (2023) is fundamentally understandable. However, there are methodological limitations, especially with regard to the lack of consideration of relevant influencing factors. These include substitution effects, important influencing factors (beyond age, gender, and BMI), and heterogeneity in price elasticities. Furthermore, the underlying database has significant weaknesses that affect the reliability of the model results. These include outdated and incomplete datasets, small sample sizes, and limited transferability to the study population. There is further criticism regarding the presentation of the model results, which suggest stronger effects and higher certainty than can actually be derived. In particular, the presentation of results in absolute values without appropriate contextualization and the inadequate consideration of limitations are in this context problematic. Given these weaknesses, the modeling results and conclusions of the study authors should be regarded as unreliable, making the study an unreliable basis for policy decisions.

4.2 Study by Schwendicke and Stolpe (2017), Germany

Full reference source:

Schwendicke, F., & Stolpe, M. (2017). Taxing sugar-sweetened beverages: Impact on overweight and obesity in Germany. *BMC Public Health*, *17*, Article 88. https://doi.org/10.1186/s12889-016-3938-4

Executive Summary

The study by Schwendicke and Stolpe (2017) examines the potential short-term effects of a hypothetical tax on sugar-sweetened beverages on overweight and obesity in Germany using a Monte Carlo simulation. The authors conclude that such a tax could lead to a significant reduction in overweight and obesity, particularly among younger people and lower income households. However, the study has methodological weaknesses that significantly limit the robustness and generalizability of the results. The underlying data, such as price elasticities from U.S. studies and outdated consumption data, do not adequately reflect the specific social and economic conditions in Germany. Furthermore, the model results rely on highly simplifying assumptions, including the linear relationship between calorie intake and weight without accounting for compensatory effects, the stable consumption response after the introduction of the tax, and the inadequate consideration of potential substitution effects. Overall, these methodological limitations significantly affect the reliability of the results and reported conclusions. Overall, from a statistical-methodological perspective, the study does not meet the required quality standards to serve as a reliable basis for evidence-based decisions.

4.2.1 Summary of the Study

The contents of the study by Schwendicke and Stolpe (2017) are summarized below.

Research Focus

The study simulates the potential short-term[37] effects of a hypothetical 20% ad-valorem excise tax on sugar-sweetened beverages in Germany. The focus is on the effects on consumption behavior and potential effects on the prevalence of overweight and obesity.

Methodology

The study considers the German population aged 15 to 79 years. The effects of the tax are estimated using a Monte Carlo simulation. Initially, changes in consumption behavior due to the tax are simulated based on price elasticities of demand and cross-price elasticities (thereafter referred to as „price elasticities“). Changes in calorie intake are derived from this and their effects on BMI are estimated, which is used to determine the prevalence of overweight and obesity.

[37] The considered effects occur within a few months after the introduction of the tax.

Results

The simulation shows that a sugar tax in Germany could result in a reduction of approximately 1.03 million cases of being overweight and 497,000 cases of obesity in the short term[38]. The largest simulated effects are seen in younger age groups and lower-income population groups. In older individuals, however, the effects are marginal or even adverse.

Conclusions of the Study Authors

The authors interpret the results as suggesting that a 20% tax on sugar-sweetened beverages in Germany would likely have positive effects on public health, especially among younger people and lower-income groups. They note that the results heavily depend on the underlying assumptions and the data on price elasticities used.

4.2.2 Evaluation of the Study

Data Foundation

The (synthetic) sample is based on a simulation of the German population aged 15 to 79 years, using official population data collected in the form of an update of the 2011 census as of May 2012 (Statistisches Bundesamt, n. d. a). While the population data was current at the time of the analysis before the study's publication in 2017, they may differ from the current German population due to significant demographic changes since then – such as migration or socioeconomic developments. This considerably limits the relevance of the results in terms of their current validity.

The data used for baseline consumption comes from a German national consumption survey called *Nationale Verzehrsstudie II* (NVS II; Max Rubner-Institut, n. d.) and are differentiated by gender, age, and income. Since the NVS II data was collected between November 2005 and January 2007, they were already about ten years old at the time of the study's publication. Therefore, potential biases in the results due to changes in consumption and market trends between the time of data collection and publication are to be expected, but they are not addressed. Furthermore, additional changes in food consumption since the publication of the study in 2017 are possible, which limits the validity of the results. The data on body mass indexes comes from the Microcensus 2013 and the KiGGS study[39]. While the Microcensus is based solely on survey data, the KiGGS study, at least in its baseline survey, includes objective measurements of body dimensions such as height and weight; later survey waves, however, mainly consist of survey data. Survey-based data can be influenced by socially desirable response behavior (e.g., incorrect reporting of body weight), which can lead to biases in the results (Schüller, 2015).

The datasets for baseline consumption and body indexes are also not fully compatible[40]. The incompatibility of the data can lead to potentially biased emphasis of individual age groups, causing deviations in the results. The study authors already pointed out the issue of incompatibility in an earlier study (Schwendicke et al., 2016), but at no point they discuss the consequences this may have on the validity of their study results.

The price elasticity data comes from U.S. studies and are differentiated only by income groups, and not further

[38]The study only provides point estimators, no confidence intervals or standard deviations.
[39]This is a study on the health of children and adolescents in Germany (Robert Koch-Institut, 2024).
[40]While the population data from the 2011 census update (as of May 2012) are classified into decade intervals (e.g., 20-29 and 30-39 years), the NVS II uses different age groups (e.g., 19-24 and 25-34 years).

by other factors (e.g., by age, gender, or consumption habits). The data is only partially transferable to Germany, as there are considerable differences in economic, cultural, and social conditions. Therefore, it is questionable whether the study results are inherently valid for Germany. Furthermore, it remains unclear whether the price elasticities also account for consumption changes in the gastronomic sector, possibly painting an incomplete picture of consumption behavior.

Model Specification

A Monte Carlo simulation (see Appendix A.3.1) is used to simulate the effects of a hypothetical tax on sugar-sweetened beverages on their consumption and BMI. The method is generally suitable for accounting for uncertainties in the input parameters. However, such a model cannot prove causal relationships. Causality is assumed as a premise and therefore needs to be upheld for each simulation step in advance. Since this is not done, the present study can only provide estimates of possible effect sizes – under the assumption that causal effects exist. The number of 100 simulation runs per group is unusually small. This does not ensure robust estimates. Additionally, it is not properly documented which distributions are applied to the varying factors in the simulation, making it difficult to trace the modeling process and independently verify and replicate the results.

The modeling of the relationships between influencing and target variables has further limitations. The change in BMI is modeled as a function of the changed calorie consumption, based on the assumption of a linear relationship between calorie intake and body weight. However, this simplification neglects the fact that overweight and obesity are influenced by complex chains of processes (Hummel et al., 2013). Neither individual metabolic differences nor the dynamic interactions between diet, exercise, and social and psychological factors are sufficiently considered.[41] The assumption of a linear relationship between energy intake and body weight strongly simplifies the actual relationships and thus presents a distorted image of the underlying mechanisms. Furthermore, the study only considers short-term effects in the form of changes in consumption behavior within a few months after the introduction of the tax, assuming that behavior changes remain stable in the long term. Potential long-term adjustments, such as consumers getting used to higher prices or changes in consumption behavior, are ignored. For a realistic modeling of BMI changes, nonlinear effects and more complex interactions would need to be considered. Furthermore, price elasticities are greatly generalized by considering them only according to two income groups (low/middle vs. high), while differences in age, gender, and baseline consumption are disregarded. This assumption overlooks the heterogeneous responses to price changes, thereby reducing the interpretability of the results.

Other relevant influencing factors are inadequately considered in the study. While fruit juices and milk are considered substitutes for sugar-sweetened beverages, other possible substitutes are not included in the simulation. Other potentially relevant factors such as regional differences, health policy frameworks, or demographic trends are also missing, which weakens the robustness of the results.

[41] The present report briefly refers to this as "compensatory effects".

Communication of Results

A negative aspect within the results is that the uncertainties are not precisely presented. Instead of providing full confidence intervals, the authors often only provide standard deviations, and even these are missing at the population level. This makes it difficult to assess the uncertainties, which are especially important for the derived measures. The results are presented as point estimates mostly in absolute numbers. Relative values, which would allow for better comparability between subgroups like age or income groups, are largely missing. A consistent presentation in both absolute and relative values would significantly improve the interpretability of the effects.

The study conducts sensitivity analyses (see Appendix A.3.2), but these are limited to isolated parameters such as cross-price elasticities and tax pass-through rates. Uncertainties arising from central model assumptions (e.g., the linear relationship between energy intake and weight change, or the assumption of long-term stable consumption) are not varied. Comprehensive sensitivity analyses that take into account key model assumptions are essential to be able to more thoroughly assess the robustness and reliability of the study results.

While the authors openly acknowledge key limitations of their study, there is a lack of a deeper discussion of these limitations. The authors point out the potentially outdated consumption data and mention the possibility of biased body indices based on survey data. They also critically examine the transfer of price elasticities from U.S. studies. The lack of differentiation of price elasticities by age and baseline consumption is mentioned, but a detailed analysis of potential biases and their impact on the results is missing. The authors discuss strongly simplifying assumptions such as the linear relationship between calorie intake and weight and the limited consideration of substitution effects. Long-term behavioral adjustments in consumption are mentioned but not discussed in terms of their potential relevance for the results. The study authors acknowledge the uncertainties of their model assumptions, but do not present them in sufficient depth. Overall, the potential impacts of all these limitations on the interpretability of the results are insufficiently addressed.

The study authors conclude from their results that the introduction of a tax on sugar-sweetened beverages in Germany may reduce overweight and obesity. Although they transparently point out that the estimated effects only hold within the framework of the assumptions made and are largely dependent on the price elasticity data used, their conclusions give the impression of causal relationships that the model cannot substantiate.

Summary of the Limitations of the Study from Schwendicke and Stolpe (2017)

Data Foundation

- **Data Sample:** Simulation of the German population (restricted to 15 to 79 years old) based on official population data from 2012, which may differ from the current German population due to significant demographic changes since then (e.g., migration).
- **Other Input Data:** Outdated consumption data (NVS II during 2005-2007) do not adequately reflect consumption and market developments and distort results; predominantly survey data (instead of measurements) on body indices, which may be biased due to self-reporting; incompatibility of consumption and body data, resulting in potentially skewed data per age group and thus deviations in results; highly generalized, potentially biased price elasticities from U.S. studies, which are only partially transferable to Germany due to country-specific differences, which makes the validity of the study results questionable.

Model Specification

- **Model Methodology:** Causality is assumed ex ante in Monte Carlo simulations during model construction, but is not proven by the model itself; the selection and distribution of parameters are not disclosed, thus limiting the replicability of the model; the small number of 100 simulation runs per group does not ensure robust estimates.
- **Influencing and Target Variables:** The simplifying assumption of a linear relationship between calorie intake and body weight contradicts the multifactorial development of overweight and obesity; only short-term effects are considered without taking into account long-term behavioral adjustments, thus limiting the interpretability of the results; failure to consider the heterogeneity of price elasticities leads to significant generalization of the results.
- **Other Influencing Factors:** Only partial consideration of possible substitutions (milk and fruit juices) and failure to consider other factors (e.g., regional differences), so the simulated effects may not accurately describe reality.

Communication of Results

- **Presentation of Results:** Missing quantification of variability at the population level and the general absence of confidence intervals make it difficult to assess random errors; the predominance of absolute numbers within the presentation complicates comparability between subgroups.
- **Model Evaluation:** There exist only isolated sensitivity analyses; lack of review of the impacts of central model assumptions make it difficult to assess model-induced uncertainty.
- **Contextualization of Study Results:** The conclusion (tax reduces overweight and obesity) claims a causal relationship that is not substantiated by the model itself; lacks a thorough discussion of the impact of the limitations on the interpretability of the results.

Overall Assessment of the Study by Schwendicke and Stolpe (2017)

The study by Schwendicke and Stolpe (2017) has significant methodological weaknesses, including an insufficient data foundation, overly simplified model assumptions, and the failure to consider relevant influencing factors. These severely limit the interpretability and reliability of the results. Outdated and inadequately matched input data, including incompatible age classifications and price elasticities with limited transferability to the target population, impair the validity of the model. Moreover, the model assumptions are based on highly simplified and unrealistic premises, such as the linear relationship between calorie intake and weight, insufficient modeling of substitution effects, and ignoring long-term behavioral adaptations. These assumptions neglect interactions and dynamics of real behavior and contradict the complexity of the factors that cause overweight and obesity. Therefore, the reliability of the results is questionable. Furthermore, the author's interpretation of the effectiveness of the tax should be avoided, as the model used is unable to demonstrate causal relationships. Overall, the evaluation shows that the study does not meet the requirements to serve as a basis for political decisions.

4.3 Study by Rogers, Cummins et al. (2023), England

Full reference source:

Rogers, N. T., Cummins, S., Forde, H., Jones, C. P., Mytton, O., Rutter, H., Sharp, S. J., Theis, D., White, M., & Adams, J. (2023). Associations between trajectories of obesity prevalence in english primary school children and the UK soft drinks industry levy: An interrupted time series analysis of surveillance data. *PLOS Medicine*, *20*(1), Article e1004160. https://doi.org/10.1371/journal.pmed.1004160

Executive Summary

The study by Rogers, Cummins et al. (2023) simulates the relationships between the events surrounding the introduction of a tax on sugar-sweetened beverages in the United Kingdom and the development of obesity prevalence in English preschool children and sixth-graders using an interrupted time series analysis. The model results show a lower obesity prevalence only in 6th grade girls when compared to a hypothetical situation without the tax. The study authors conclude that the tax could contribute to the reduction of obesity prevalence in older elementary school children, but that the tax alone is not sufficient for a widespread reduction in obesity in children. However, the reliability of the model results is significantly limited, particularly due to the non-representative sample and the failure to consider other relevant factors, which prevents the effects to be clearly attributed to the tax. This undermines the robustness of the already highly generalized conclusions. Overall, the study does not meet the required quality standards from a statistical-methodological perspective to serve as a reliable basis for evidence-based decisions.

4.3.1 Summary of the Study

The contents of the study by Rogers, Cummins et al. (2023) are summarized below.

Research Focus

The study models the relationships between the events surrounding the announcement of the UK's *Soft Drinks Industry Levy* (SDIL)[42] and the development of obesity prevalence in English children, with a particular focus on possible gender differences and regional socioeconomic disparities.

Methodology

The study focuses on children who attended preschool (ages four to five) or sixth grade (ages ten to eleven) in public schools in England between September 2013 and November 2019. The analysis relies on the children's height and weight data, from which the prevalence of overweight and obesity is determined. Socioeconomic status is determined based on the location of the school. In the main analysis, the effect of the tax announcement (March 2016) is investigated using an *interrupted time series analysis*, comparing the real data to a simulated

[42] Manufacturers are subject to a tiered excise tax on packaged, sugar-sweetened beverages, depending on the sugar content: Beverages with less than 5 g of sugar per 100 ml are exempt, those with a sugar content between 5 and 8 g per 100 ml are taxed at 0.18£ per liter, and those with 8 g per 100 ml or more are taxed at 0.24£ per liter (Sasse & Metcalfe, 2022).

trend without the tax announcement (= counterfactual scenario).[43] The analysis is conducted both overall and separately for different socioeconomic groups. The time series are analyzed using an *ARIMA*[44] model. In addition to the main analysis, various sensitivity analyses (see Appendix A.3.2) are carried out that consider on the one hand different definitions of the intervention (i.e., interruptions in the time series), namely (1) the reformulation of sugar-sweetened beverages starting in November 2016[45] and (2) the actual introduction of the tax in April 2018, and on the other hand, (3) extend modeling obesity prevalence to the prevalence of overweight and obesity.

Results

According to the modeling results, the obesity prevalence in girls in the 6th grade is 1.6 percentage points lower at the end of the study period, 95 % CI [1.1; 2.1], compared to the simulated trend without the tax, which corresponds to a relative reduction in obesity prevalence[46] of 8.0 %, 95 % CI [5.4; 10.5]. In socioeconomically disadvantaged areas, the model shows a stronger effect (2.4 percentage points in the most disadvantaged area, 95 % CI [1.6; 3.2]). In some subgroups, such as boys in the 6th grade with schools located in socioeconomically privileged areas, the modeling shows a higher obesity prevalence of 1.6 percentage points, 95 % CI [0.7; 2.5], or 10.1 %, 95 % CI [4.3; 15.9]. No effects are found for preschool children. Similar results are seen when the intervention is defined as the point of change in the sugar content for sugar-sweetened beverages (Sensitivity Analysis 1) or when considering the combined prevalence of overweight and obesity (Sensitivity Analysis 3). With the intervention defined as the point of tax introduction (Sensitivity Analysis 2), no significant differences in obesity prevalence are observed.

Conclusions of the Study Authors

Based on their modeling results, the study authors conclude that SDIL can help reduce obesity in older elementary school children, and health inequalities among children in general. Furthermore, they advocate for additional measures to reduce obesity prevalence in children, especially in boys and younger age groups.

4.3.2 Evaluation of the Study

The study by Rogers, Cummins et al. (2023) is evaluated below based on the evaluation criteria presented in Chapter 3.

Data Foundation

The study is based on cross-sectional data on the height and weight of preschool children and of 6th grade students. The data is collected annually only for these age groups as part of the *National Child Measurement Program* (short: *NCMP*; Office for Health Improvement and Disparities, 2024). The measurements are generally objective and cover almost all state schools in England.

The participation rates are high (about 90 %), as parents must actively object to exclude their children from the measurements. An indication of a non-representative sample arises from the observation that overweight and

[43] The timing of the tax announcement marks the beginning of the intervention, from which point the actual data is compared with a counterfactual scenario. This point can therefore be seen as an “interruption” of the time series.

[44] ARIMA stands for “autoregressive integrated moving average”.

[45] Already after the SDIL announcement, many manufacturers adjusted their formulations to stay below the sugar content thresholds that require (higher) fees (Dickson et al., 2023). The number of products that modified the recipe increased at the end of 2016 (Sasse & Metcalfe, 2022).

[46] In detail, this is not a reduction in obesity prevalence after the announcement of the SDIL, but a deceleration of the increase in obesity prevalence within the overall increasing trend of childhood obesity.

obese girls tend to be more frequently excluded from the program. According to the NCMP guide, this selection could lead to biases regarding girls (Hancock & Copley, 2016). Since the results of Rogers, Cummins et al. (2023) show a significant reduction in obesity prevalence compared to the counterfactual scenario only for girls in 6th grade, it must be carefully examined whether an underrepresentation of overweight and obese girls in the sample is responsible for this effect. Since different children are measured each year as part of the NCMP, the data does not provide information about individual developments over time, but instead represents cross-sectional distributions of body indices at the respective measurement time. Additionally, the data for preschoolers was collected at different times during the school year. The published figures show the data points of the analyzed time series at monthly intervals. However, more detailed information on the monthly composition of the sample (which the obesity and overweight prevalence were calculated on) is missing. Therefore, systematic differences over time cannot be excluded, which calls into question the randomness and unrestricted comparability of the data points.

The classification for overweight and obesity based on the NCMP data follows the standards of the National Health Service (National Health Service, 2023). Overweight and obesity are not categorized by a fixed BMI value, but instead by age-dependent distributions according to Cole and Lobstein (2012), which allows consideration for the natural variability of BMI across different age groups. The reference to consistent reference values supports comparability. However, it is worth questioning whether the underlying reference data from 1990 is still relevant, as changes in population structure and lifestyle since then may have influenced BMI distribution.

The estimation of socioeconomic status based on the school location does not consider individual social factors and should therefore be only as a rough classification with a high degree of uncertainty.

Model Specification

Interrupted time series analyses (see Appendix A.2.1) are generally suitable for identifying temporal changes (influencing variable: time) of a target variable (development of obesity prevalence) before and after an intervention. The direction of influence on the target variables is assumed beforehand in a time series analysis and is not proven by the model. Since the target variables in reality are often influenced by many factors simultaneously, determining the isolated effect of the intervention is difficult. For example, obesity prevalence may be influenced not only by the introduction of the SDIL, but by other factors such as education campaigns. Interrupted time series analyses can hardly control for such influences and their interactions, which makes isolating the effect of the SDIL difficult. Such influencing factors are not considered in the model by Rogers, Cummins et al. (2023). Their model only considers age, gender, and the months around the summer holidays[47] as predictors.

Additionally, in a time series analysis, the calculation of the counterfactual scenario (simulated trend without tax) is highly sensitive to the chosen point used for modeling the trend. To simulate the possible trend without the tax, the historically observed data is used up to the intervention point. However, if the (unknown) intervention point is incorrectly modeled, this can lead to distorted estimates of the effects. To assess such a possible misspecification, the study authors conduct sensitivity analyses with two additional intervention points, allowing for a better assessment of the actual situation.

[47] For preschool children, the months of September, October, June, and February were considered, and for 6th-grade school children, the months of September and July were used. These were identified as relevant for obesity prevalence in a preliminary analysis.

Jandoc et al. (2015) recommend uniform methods and reporting standards for interrupted time series analyses, which the study by Rogers, Cummins et al. (2023) largely follows. However, it is possible to recognize weaknesses and gaps in the methodological specification. (1) After the announcement of the SDIL, the counterfactual scenario assumes a simplified linear continuation of the obesity prevalence trend. However, it is unlikely that the development of obesity prevalence is strictly linear in reality, so the derived results likely contain uncertainties. (2) Essential information[48] about the ARIMA model and the verification of the required assumptions is missing. The missing information implies that Rogers, Cummins et al. (2023) either incompletely document the model specification or mistakenly refer to a linear regression model as an ARIMA model, which is generally the most suitable method for modeling seasonality, autocorrelation, and noise. This limits the replicability of the underlying model. Without comprehensively examining the assumptions, the assessment of the validity of the results is limited. However, estimation of the model parameters is generally done using standard methods tailored to the data base. (3) Similarly, Rogers, Cummins et al. (2023) do not provide explicit information on whether the number of underlying data points is sufficient, even though Jandoc et al. (2015) emphasize the importance of a sufficient number of data points for the robustness of time series analyses. However, the data points shown in the study figures suggest that the database for conducting the analyses can be regarded as sufficient in principle.

Communication of Results

All results are transparently provided along with confidence intervals. However, it should be noted that confidence intervals reflect only random sampling errors, and not the non-random errors resulting from the non-representative sample (underrepresentation of overweight and obese girls). Therefore, there is limited interpretability of the provided confidence intervals. In addition, the simulated differences in obesity prevalence between the observed and simulated trends are presented only as absolute or relative differences in percentages or percentage points. There is no contextualization (e.g., converting to absolute numbers to state the number of obese children in the scenarios with and without tax; or also natural frequencies). This complicates the interpretation of the results and reduces the transparency of the analysis. Information on the calculated trend parameters, i.e., the rates of change of the real and counterfactual trends, is missing, making it difficult to verify the plausibility of the results. Furthermore, contrary to the recommendation of Greenland et al. (2016), information on p-values and effect sizes is missing, which would allow for a more informed interpretation of the statistical relevance of the differences between the real and simulated trends.

The study authors discuss several limitations of their study, including the selective participation of obese girls, without plausibly explaining why this (as they assumed) should lead to an underestimation of the effects. Additionally, they list the rough estimation of socioeconomic status, the lack of data after the introduction of the SDIL to assess long-term effects, and the insufficient examination of seasonal variations as limitations. The sensitivity of the simulated trend without the tax is also discussed: Depending on the chosen intervention point, the estimated effect of the tax could be significantly skewed. To reduce this risk, the study conducts sensitivity analyses with different intervention points. The study authors also correctly report that their modeling result is merely a (correlational) relationship between the announcement of the SDIL and a reduction in obesity cases among 10- to 11-year-old girls.

[48] Missing values for autoregressive terms, difference terms to remove trends or non-stationarity, error terms as moving-average components, and complete information about shift parameters and their investigation such as the treatment of non-stationarity.

Based on their modeling results, the study authors conclude that the SDIL can contribute to reducing obesity among older primary school children, but this assumption is not sufficiently supported by the model. The effect is based solely on the results for 6th grade girls, whose data is likely biased due to the underrepresentation of overweight and obese girls in the sample. Drawing such an effect is overgeneralized, especially since no comparable effect was found for boys. The study authors therefore also conclude that the tax alone is not enough to reduce obesity in children – particularly in boys and younger age groups – and that additional measures are required. Since only preschool children and 6th graders were studied, an analysis of all age groups could provide more detailed insights. A call for additional measures can therefore be only partially derived from the results.

Furthermore, the study authors conclude that the SDIL may contribute to reducing health inequities related to obesity among children[49]. However, this conclusion is simplified, as it is based on a rough classification of socioeconomic status without considering individual differences. Additionally, it relies largely on presumably biased effects in the sample of girls. Depending on gender and socioeconomic status, the model shows both significant and non-significant effects in varying directions (see Table 2 in the study by Rogers, Cummins et al., 2023). The proposed potential effect of improving health equity remains highly hypothetical and undoubtedly requires more specific analyses to provide reliable conclusions, particularly with regard to the design of political measures.

[49] According to Rogers, Cummins et al. (2023), the prevalence of obesity is higher in socioeconomically disadvantaged areas than in privileged ones. A reduction in health inequity means that the obesity prevalence in disadvantaged areas rises less sharply after the tax than in privileged ones, resulting in a general leveling between the different socioeconomic groups.

Summary of the Limitations of the Study by Rogers, Cummins et al. (2023)

Data Foundation

- **Data Sample:** Structurally different sample of the subgroup of girls (underrepresentation of overweight and obese girls), leading to biased results and specifically calling into question the simulated effect for sixth-grade girls.
- **Other Input Data:** Body indices as a series of cross-sectional data potentially not random, meaning systematic differences over time cannot be excluded; only a rough (regional) classification of socioeconomic status, conclusions based on it are uncertain.

Model Specification

- **Model Methodology:** Causality is assumed beforehand in time-series analyses during the construction of the model, but not proven by the model itself; general sensitivity of the time-series analyses in calculating the simulated trend (scenario without tax) to the intervention point with the risk of biased effect estimates; insufficient details on model specification (e.g., ARIMA vs. linear regression; missing tests for necessary assumptions), thus limited replicability.
- **Influence and Target Variables:** Simplified assumption of a linearly continued trend of obesity prevalence after the announcement of the SDIL for the counterfactual scenario, which may not align with the actual development and likely worsens the model fit.
- **Other Influencing Factors:** Failure to account for relevant factors (e.g., educational campaigns), so that the effects of SDIL do not necessarily exist in reality.

Communication of Results

- **Presentation of Results:** Despite transparent reporting, the confidence intervals have limited explanatory power, as they cannot capture uncertainty caused by the non-representative sample; the absence of p-values and effect sizes prevents a thorough assessment of statistical significance; lack of contextualization complicates result interpretation; missing methodological details on trend parameters put into question the plausibility of results.
- **Model Evaluation:** Despite conducting sensitivity analyses on intervention points, results are uncertain due to the high sensitivity of the simulated trends to these points.
- **Contextualization of Study Results:** The conclusion (SDIL can contribute to reducing obesity or health inequalities in older schoolchildren) is insufficiently supported by the model (effect is solely or largely due to the simulated result for the subgroup of sixth-grade girls, which is likely skewed); the call for additional measures to reduce obesity prevalence is only partially derivable from the results.

Overall Assessment of the Study by Rogers, Cummins et al. (2023)

The reliability of the modeling results by Rogers, Cummins et al. (2023) is significantly limited due to the non-representative sample and the failure to account for additional relevant factors. The lower participation of overweight and obese girls in the measurements leads to potential bias in the modeling results, which the authors mention but do not sufficiently discuss. Because of the failure to account for relevant factors (e.g., parallel educational campaigns), the simulated effects cannot be clearly attributed to the tax. These weaknesses undermine the reliability of the conclusions, so the study should not be used as a reliable basis for policy decisions.

4.4 Study by Cobiac et al. (2024), England

Full reference source:

Cobiac, L. J., Rogers, N. T., Adams, J., Cummins, S., Smith, R., Mytton, O., White, M., & Scarborough, P. (2024). Impact of the UK soft drinks industry levy on health and health inequalities in children and adolescents in England: An interrupted time series analysis and population health modelling study. *PLOS Medicine*, *21*(3), Article e1004371. https://doi.org/10.1371/journal.pmed.1004371

Executive Summary

The study by Cobiac et al. (2024) models the impact of a sugar-sweetened beverage tax introduced in the United Kingdom in April 2018 on sugar consumption from beverages in English households and the associated health impacts on children up to 17 years old, taking into account socioeconomic disadvantage. The model is based on an interrupted time series analysis to capture changes in sugar consumption. Potential effects of the consumption change on obesity and health-related aspects are modeled using two cohort simulations. The modeling results show a reduction in sugar consumption in English households, particularly in socioeconomically disadvantaged regions. Additionally, the model predicts a reduction in the prevalence of obesity, an improvement in dental health and quality of life, and a negligible change in life expectancy in children. The results should be critically evaluated as the study has fundamental weaknesses in its data foundation: The sales data used do not adequately reflect the actual consumption of the sample studied. Furthermore, simplifying model assumptions and the neglect of relevant influencing factors further undermine the reliability of the results. Overall, from a statistical-methodological perspective, the study does not meet the required quality standards to serve as a reliable basis for evidence-based decisions.

4.4.1 Summary of the Study

The contents of the study by Cobiac et al. (2024) are summarized below.

Research Focus

The study models the potential effects of the sugar-sweetened beverage tax as part of the British SDIL announced in March 2016 and implemented in April 2018 on sugar consumption from beverages in English households, as well as the resulting medium-term and long-term health impacts on children, taking into consideration the degree of regional socioeconomic disadvantage.

Methodology

Based on data from weekly household purchases, an interrupted time series analysis is used to simulate how the sugar amount acquired through the purchase of various beverages – not exclusively sweetened with sugar – changed from March 2014 to November 2019 compared to a counterfactual scenario without the tax. On this basis – assuming a stable purchasing trend over a period of ten years – medium-term effects on BMI, dental cavities and quality of life (QALYs)[50] as well as long-term effects on the life expectancy of English children between the ages of zero and 17 are modeled using two different cohort simulations. The analyses are conducted for the individual as well as the combination of different socioeconomic groups.

Results

The simulation results show that, after the introduction of the SDIL, the amount of sugar acquired through beverage purchases decreased by 15 grams per household per week, 95 % CI [10.3; 19.7]. According to the model predictions, this reduction would lead to 3,600 fewer cases of dental cases, 95 % CI [946; 6,330], and 64,100 fewer cases of overweight and obesity, 95 % CI [54,400; 73,400] by 2025, resulting in a total gain of 19,500 QALYs, 95 % CI [14,800; 24,600]. The predicted change in life expectancy is marginal[51] and is therefore negligible.

Conclusions of the Study Authors

The study authors conclude from the reduction in sugar acquired through beverage purchases, that the SDIL has led to a reduction in the sugar content of beverages, which could have a positive impact on health aspects in English children.

4.4.2 Evaluation of the Study

The study by Cobiac et al. (2024) is evaluated below based on the assessment criteria presented in Chapter 3.

Data Foundation

Population data of the entire child and adolescent population (0 to 17 years) in England from 2015, as well as population projections for modeling birth cohorts after 2015, are used for the forecast of the *medium-term* health and health-related impacts of the SDIL on children. Population data from the 2015 birth cohort in England is used for the forecast of the *long-term* health-related impacts. All of this data comes from the Office for National Statistics. Since the data is current official population data, which represents the entire target population, a fundamental structural match with the target population can be assumed.

The data on beverage purchases[52] is derived from weekly cross-sectional data on household purchases, covering an average of about 17,000 English households. Data providers are households that provide their weekly purchase data to a market research company for a small annual fee. The data was collected based on self-reports. Due to potentially socially desirable responses or recall bias, these are generally prone to be skewed. To ensure representativeness, Cobiac et al. (2024) assign weights to different factors from the market research company, but they do not explain the process in detail. To calculate the per capita change in sugar purchase through beverages from household-level data, an additional source for the average household size (2.4 persons) is used. It is unclear whether the match between the average household size in the sample and that of the population was verified. Furthermore, it is problematic to use data from an average household member when studying children, as this does not adequately reflect the specific purchasing behavior of the target group[53]. Additionally, beverage purchases were calculated solely based on retail data, without considering service industry data, which leads to

[50] Quality Adjusted Life Years (QALYs) are calculated from the predicted remaining life span, multiplied by a utility factor that is between zero for the worst conceivable and one for the best possible quality of life.

[51] Depending on the degree of socioeconomic disadvantage, it ranges from a few days to a month.

[52] This is specifically purchase data, not consumption data. The associated study limitations are explained in the following section in *Model Specification*.

[53] The consequences for the validity of the study results are explained in the *Result Communication* section.

biased results. Cobiac et al. (2024) assume, referencing Cornelsen et al. (2019), that the share of spending on purchases outside the household is about 10-12 %. Other sources estimate this share significantly higher, about 45 % (Statista, 2022). These shares may vary by region and socioeconomic status (Law et al., 2022), which is also noted by Cobiac et al. (2024).

The assignment of the degree of socioeconomic disadvantage of a household is based on the households' postal codes according to the *Index of Multiple Deprivation* (Smith et al., 2015), allowing for rough regional estimates but ignoring individual factors not directly related to geographic location. Based on this, households in the model are categorized into five equal groups, stratified by the degree of disadvantage. However, this does not achieve an adequate representation and separation of social disparities.

The model parameters indicating health come from various data sources and rely on different data collection methods. Some is cross-sectional data from government-conducted surveys on body indices and dental health, collected through self-reports and partially supplemented by objective measurements (National Health Service, 2011, 2015, 2016). The majority of health parameters are based on data from England (analogous to the consumption data), with only dental cases including data points from Wales and Northern Ireland, meaning there is no complete geographic coverage. Considerable deviations from purely English data are not expected, as Wales and Northern Ireland are also part of the United Kingdom and are subject to the SDIL in the same way. Data on the quality of life and life expectancy come from meta-studies, whose validity strongly depends on the methodology of the included studies. The applicability in the specific study context and the quality of the original data remain unclear. The parameters for estimating the quality-adjusted life years based on dental health and being overweight were originally developed and validated for adults, but there is limited evidence for their appropriateness in children (Brown et al., 2018).

Model Specification

To exclude effects from Brexit or the COVID-19 pandemic, the period from March 2014 to November 2019 was chosen to estimate the change of the amount of acquired sugar through beverage purchases. This increases the interpretability of the results. Three different modeling methods are used to derive the results:

(1) To estimate the change in the amount of sugar purchased through beverages in a household, an interrupted time series analysis (see Appendix A.2.1) is conducted with two intervention points (announcement and introduction of the SDIL), considering the average monthly temperatures and the (post) Christmas period[54]. As described in section 4.3.2, interrupted time series analyses are generally sensitive to the choice of the intervention point. The direction of effect from the independent to the dependent variables is assumed beforehand in time series analyses and is not proven by the model. If control variables are not adequately considered, the observed effects of the SDIL may not necessarily exist in the form presented by the model.

(2) To estimate the prevalence of overweight (and obesity) as well as dental health in children (middle-term effects), a cohort simulation (see Appendix A.1.1) is conducted over a period of ten years (2015 to 2025), assuming the change in sugar purchased through beverages as a central influencing factor. Subsequently,

[54]These control variables account for seasonal trends. The relevance of December and January months was identified in a preliminary analysis.

the cohort simulation is expanded to forecast the QALYs of children, with obesity prevalence and dental health as influencing factors. The direction of effect from the independent to the dependent variables is assumed beforehand in cohort simulations and is not proven by the model. If control variables are not adequately considered, the observed effects of the SDIL may not necessarily exist in the form presented by the model.

(3) To forecast life expectancy (long-term effects), a cohort simulation is conducted over a period of 100 years, considering BMI and dental health as influencing factors.

In all three methods, the actual situation of the tax introduction is compared with a counterfactual scenario without SDIL.[55] It is assumed that the observed tax effect will remain unchanged in the long term, without consideration of possible dynamic effects such as consumer behavior adjustments, producer price strategies, or changes in the market structure. The neglect of this dynamic represents a significant simplification of the model compared to the reality and may have a major impact on the reliability of the results, especially with a modeling period spanning several decades.

The study overly simplifies by equating the change in purchasing behavior with a change in consumption behavior: (1) It is assumed that the sugar purchased through beverages is completely consumed, which is not necessarily the case. (2) The lack of differentiation by age and gender completely disregards the heterogeneity of consumption. Therefore, it cannot be assumed that the average household sugar consumption from beverage consumption is representative of children's consumption.

Additionally, the estimated sugar consumption based on the change purchasing behavior is converted into calories and linearly translated into BMI changes using energy balance equations, without considering compensatory effects such as changes in physical activity or caloric substitution from other foods. Moreover, a fixed dose-response relationship[56] between sugar and dental cavities is assumed in the prediction, without considering additional factors like dental hygiene or fluoride. Overall, a realistic depiction of the actual relationships is not drawn, which limits the reliability of the results.

A central influencing factor considered by the study is the degree of regional socioeconomic disadvantage. However, the modeling neglects several other relevant influencing factors. The study authors mention that sugar consumption from purchased beverages had already decreased before the tax announcement, possibly due to other concurrent measures like awareness campaigns. However, this potential influencing factor is not considered in the model. Substitution effects are also completely neglected. Therefore, the simulated effects can hardly be solely attributed to the tax, as presented by the model.

[55] For this, the trend before the SDIL is extrapolated to estimate what the values would have been without the tax. The estimated changes due to the tax strongly depend on the chosen point in time used for determining the trend. To control for general trends in household purchases that are not influenced by the SDIL, the authors use purchases of toilet items such as shampoo and soap as control variables. These are considered largely unaffected by seasonal or socioeconomic factors. However, they exhibit different purchasing behavior than the acquisition of (sugar-sweetened) beverages, which may limit their suitability as a control variable in this context.

[56] The dose-response relationship describes how a change in the amount (dose) of a factor affects the strength of the effect (Bernabé et al., 2016).

Communication of Results

All presented results (differentiated by degree of disadvantage) include confidence intervals, making the uncertainty due to possible sampling error clear. However, further statistical classification of the effects, such as by calculating *p*-values or effect sizes, is not provided. This excludes a thorough statistical classification. The modeling results are presented in absolute values (e.g., reduction of 3,600 cases of dental cases, 95 % CI [946; 6,330]), but are not further contextualized. For complete interpretation of the results, information on the underlying base (e.g., total number of caries cases in the studied population) is missing. Additional information on natural and relative frequencies could have facilitated the interpretation.

In validating their results, the study authors point out that the simulated reduction in sugar intake from beverages in English households of 15.0 g per household per week[57], 95% CI [10.3; 19.7], is nearly double the reduction simulated by Rogers, Pell et al. (2023) for the UK. Both studies use comparable methods but rely on different datasets. Cobiac et al. (2024) attribute this discrepancy to slight differences in the time series data on sugar consumption before the introduction of the SDIL. This demonstrates how sensitive (long-term) results are to minimal changes in the input data. Cobiac et al. (2024) themselves highlight that even slight deviations from the initial reduction in sugar purchases of ±0.1 g per month can influence the counterfactual estimates at the end of the simulation period (November 2019) by ±4.4 g of sugar. Therefore, the simulated results are subject to a certain degree of uncertainty and should generally be considered with caution. Cobiac et al. (2024) do not conduct systematic sensitivity analyses to better assess the impact of this sensitivity on the simulated effects.

Furthermore, the interpretability of the results is additionally skewed by the previously mentioned lack of differentiation in sugar consumption between children and adults, particularly in combination with the fact that purchase data, rather than actual consumption data, is used. According to the calculations of Cobiac et al. (2024), based on purchase data, 341.6 g of sugar per week per household (equivalent to 142.3 g per household member) were purchased through beverages just before the announcement of the SDIL in March 2016. In comparison, an actual consumption study, the *National Diet and Nutrition Survey 2016* (Public Health England, 2018), shows that ten-year-old children consume an average of 80.1 g of sugar per week.[58] This value includes, unlike the data source of Cobiac et al. (2024), beverages consumed outside the home, and is still significantly lower than the per capita sugar purchase from beverages calculated by Cobiac et al. (2024). Since the effects of the SDIL simulated by Cobiac et al. (2024) are not based on actual consumption, they may be significantly overestimated and should be critically questioned.

The study authors address various assumptions, uncertainties, and limitations of their study. For example, they mention that co-interventions, such as awareness campaigns, make it difficult to clearly isolate the effect of the SDIL. They also discuss that the counterfactual scenario is based on the assumption that purchasing trends would have continued linearly without the intervention, and that the effects of the SDIL would remain constant in the long term – assumptions that must be critically examined. They also mention that calculating the average consumption per person disregards individual differences within households. They also state that

[57] A reduction of 15.0 g of sugar per week per household with an average of 2.4 people corresponds to a reduction of 0.89 g of sugar per person per day, which equals an energy savings of less than 3 kcal per day (own calculation based on values from Cobiac et al., 2024). This reduction appears extremely small.

[58] According to Public Health England (2018), they consume 52 g of sugar per day, i.e., 364 g per week. Of that, 22 % is attributable to beverages. Thus, 80.1 g of sugar per week comes from beverages.

compensatory behavioral changes, such as adjustments in physical activity, were also not considered. They justify this by assuming that such changes had not occurred within the three-year observation period after the announcement of the SDIL. They also note that the estimates of quality-adjusted life years and life expectancy should be interpreted with caution due to significant fluctuations and uncertainties in the parameters derived from meta-studies.

Overall, the results are interpreted in a highly generalized way: (1) While the study authors correctly note in the results section that the changes in life expectancy are minor (without quantifying how small these effects actually are), the summary states that the SDIL will lead to long-term improvements in life expectancy. This conclusion, however, is not justifiable, as the effects, depending on the degree of socioeconomic disadvantage, only range from a few days to at most one month, and thus should be considered small. The statement of the study authors therefore represents a significant overestimation of the simulated effects, especially since the uncertainty of the results, which must be taken into account especially in the case of long-term effects, is completely disregarded. (2) The study authors correctly point out that the health outcomes, such as reduced BMI values or fewer dental cavity cases, are mainly correlatively associated with the introduction of SDIL, as other relevant influencing factors were not taken into account in the model. Nevertheless, they draw a causal conclusion by claiming that SDIL leads to a reduction in overweight, obesity and dental cavity as well as to an improvement in quality of life.

Summary of the Limitations of the Study by Cobiac et al. (2024)

Data Foundation

- **Data Sample:** There are no limitations regarding the representativeness of the sample due to the use of official population data representing the entire target population.
- **Other Input Data:** Beverage purchase data for an average person (especially based on self-reports and without accounting for the service industry) do not accurately reflect the specific purchasing behavior of the sample and skews results; possible incompatibility of data sources (household purchases and household size) could lead to potential deviations in the results; health data partially based on self-reports and meta-studies with unclear quality of original data and unproven suitability for the study context, making the results potentially uncertain; only rough (regional) classification of socioeconomic status, meaning conclusions regarding this are also uncertain.

Model Specification

- **Model Methodology:** Causality is assumed beforehand in time series analyses and cohort simulations during model construction, but is not proven by the model itself; general sensitivity of time series analyses when calculating the simulated trend (scenario without tax) to the intervention point with the risk of biased effect estimates.
- **Influencing and Target Variables:** Numerous simplifying assumptions (long-term constant tax effects, equating consumption with purchasing behavior, neglecting heterogeneity in consumption behavior, linear relationship between calorie intake and weight neglecting compensation effects, fixed dose-response relationship between sugar and caries neglecting individual oral hygiene) do not accurately reflect real phenomena and their relationships.
- **Other Influencing Factors:** Failure to consider relevant factors (e.g., substitution, awareness campaigns) means that the simulated effects may not accurately reflect the real relationships.

Communication of Results

- **Presentation of Results:** The absence of p-values and effect sizes prevents a proper statistical interpretation of significance; the lack of contextualization makes it more difficult to interpret the results.
- **Model Evaluation:** Discrepancies in results compared to other studies (sugar consumption and change) indicate biases in the results and an overestimation of the effects; the sensitivity of the results to minimal changes in the input data leads to uncertainties that are not assessed through sensitivity analyses.
- **Contextualization of Study Results:** Neglect of uncertainty and overestimation of effects (life expectancy); the causal conclusion (SDIL leads to reductions in overweight, obesity, and dental cavity, as well as improved quality of life) is not supported by the model.

Overall Assessment of the Study by Cobiac et al. (2024)

The reliability of the study results by Cobiac et al. (2024) must be critically questioned, particularly due to the insufficient data basis, numerous simplifying model assumptions, and the failure to account for relevant influencing factors. Particularly problematic is that the simulated sugar consumption of adolescents is based on household-level sales data, which does not realistically reflect actual intake. Health-related effects are modeled on this fragile foundation, which excludes reliable conclusions. The overly simplified assumptions about the relationship between the tax and its effects, such as the linear relationship between calorie intake and weight without considering caloric compensation or other influencing factors on body weight, do not reflect reality. Furthermore, it is likely that the simulated effects do not exist in the form presented but are partially attributable to other influencing factors, such as awareness campaigns. Validations also suggest that the effects simulated by Cobiac et al. (2024) were overestimated. Given these weaknesses, the modeling results and conclusions of the study authors should be considered unreliable, and therefore, the study does not provide a solid foundation for policy decisions.

4.5 Study by Gračner et al. (2022), Mexico

Full reference source:

Gračner, T., Marquez-Padilla, F., & Hernandez-Cortes, D. (2022). Changes in weight-related outcomes among adolescents following consumer price increases of taxed sugar-sweetened beverages. *JAMA Pediatrics*, *176*(2), 150–158. https://doi.org/10.1001/jamapediatrics.2021.5044

Executive Summary

The study by Gračner et al. (2022) examines the relationship between price changes of sugar-sweetened beverages as a result of the tax introduced in Mexico and body weight related parameters among Mexican adolescents living in cities using multivariate regression models. The modeling results show that a 10 % increase in the price of sugar-sweetened beverages among adolescent girls is associated with a 3 % relative decrease in their prevalence of overweight and obesity. The observed weight losses are small and occur predominantly among girls with higher initial weight who live in cities where the price increases were higher than average. No comparable effects were found for boys. The study authors conclude that large increases in price could be associated with observable changes in weight-related parameters. However, in view of the reported results and numerous study limitations, this conclusion appears to be highly generalized. In particular, the selection of the sample, which only includes adolescents living in cities and with health insurance from a specific insurance company, significantly limits the transferability of the results to the entire (adolescent) population in Mexico. Overall, from a statistical and methodological perspective, the study does not meet the necessary quality standards to serve as a reliable basis for evidence-based decisions.

4.5.1 Summary of the Study

The contents of the study by Gračner et al. (2022) are summarized below.

Research Focus

The study investigates the relationship between price changes[59] of sugary drinks following the specific excise tax of 1 Mex$ per liter introduced in Mexico in January 2014 and various weight-related parameters, including BMI and the risk and prevalence of overweight and obesity[60] in adolescents in Mexican cities.

Methodology

The study examines 12,654 adolescents from 39 Mexican cities, born between 1999 and 2002. Based on longitudinal data on their body weight and height, the prevalence of overweight and obesity in the study group is simulated for the period 2012 to 2017 using multivariate regression models. The analysis also takes into account the extent of price changes for taxed drinks, with cities divided into three categories (price increase < 5 %, 5-10 %, and > 10 %). Sensitivity analyses are also performed to check the robustness of the results.

[59] Price changes varied by city.
[60] Overweight and obesity are not further differentiated in this study.

Results

The modeling results show that a 10 % price increase for sugary drinks[61] among girls within two years after the price change was associated with an absolute decrease in the prevalence of overweight and obesity by 1.3 percentage points, 95 % CI [−2.1; −0.38], $p = 0.006$, and a relative decrease of 3.0 %[62]. In girls with higher initial weight, a reduction in their BMI is simulated, corresponding to a weight loss of about 0.35 kg[63]. In cities where price increases due to the tax were above average, i.e., more than 10 %[64], larger effects were simulated for the girls' results. No such relationships were found for boys.

Conclusions of the Study Authors

The study authors assess the simulated weight losses, which primarily occurred in girls with higher initial weight and living in cities with price increases of more than 10 %, as small. They conclude that large price increases might be associated with observable changes in weight-related parameters.

4.5.2 Evaluation of the Study

The study by Gračner et al. (2022) is evaluated based on the assessment criteria presented in Chapter 3.

Data Foundation

The sample of young people born between 1999 and 2002 was ten to 18 years old during the observation period (2012-2017). To ensure compatibility with the price data, only health data from adolescents from 39 Mexican cities[65] is included, but not from rural areas. There are typically systematic differences between urban and rural regions, for example in socioeconomic status or access to healthcare facilities. The data on adolescents' body measurements come from the public health institute *Instituto Mexicano del Seguro Social* (short: *IMSS*) and include adolescents who had health insurance through the IMSS and who received annual[66] outpatient health treatments within the same city. Their body measurements were also recorded as part of these treatments. For each adolescent, six (annually averaged) data points on body weight and height were available for the analysis. With the help of such longitudinal data, it is also possible to observe individual changes over time. However, the representativeness of this data is limited: According to the authors of the study, the IMSS insures employees in formal employment and their families[67]. As a result, young people from households with more stable income and education levels could be overrepresented. In addition, there may be systematic differences between adolescents who were screened annually at the clinic (as in the sample) and those who did not meet this criterion (not in the sample). According to the study authors, the sample studied is also younger[68], slightly heavier[69] and more often female[70] than the average Mexican adolescent. The study results are therefore not fully transferable to the entire Mexican adolescent population, especially not to adolescents living in rural areas.

[61] This corresponds to the average price increase for sugary drinks in Mexico.
[62] No corresponding confidence interval given.
[63] No corresponding confidence interval given.
[64] The average price increase in these cities is given as 16.6 %.
[65] In detail: urban regions within a radius of 30 kilometers from the city center.
[66] at least once and a maximum of ten times per calendar year; on average 3.59 treatments per year
[67] The IMSS insures more than 50 % of the Mexican population in total.
[68] The average age of the adolescents studied is given as 11.38 years and presumably refers to the starting point of the study in 2012.
[69] According to the study authors, the risk and prevalence of overweight and obesity is higher in the IMSS sample, which is presumably due to the fact that heavier people are more likely to attend a clinic
[70] According to the study authors, this reflects the fact that adolescent girls tend to use health services more frequently than boys.

The calculated inflation-adjusted annual average prices for sugar-sweetened beverages are based on price data from the National Institute of Statistics and Geography for the period 2011-2016. This data includes monthly price data for products sold in stores in Mexican cities, but does not include information on sales in food service establishments. Only *carbonated* sugar-sweetened beverages were included in the analysis, as data for non-carbonated sugar-sweetened beverages was not fully available for all cities analyzed. According to the authors of the study, however, young people mainly consume carbonated, sugar-sweetened drinks. In addition, prices would have changed significantly after the tax increase, especially for carbonated sugar-sweetened beverages. Due to the fact that prices in catering outlets and non-carbonated beverages are not taken into account, deviations between the price changes used and the actual price changes for sugar-sweetened beverages cannot be ruled out. The cities are divided into three groups based on the observed price changes: low (< 5 %; nine cities[71]; 1,499 young people), medium (5-10 %; seven cities; 6,986 young people) and high (> 10 %; 23 cities; 4,169 young people) price changes.

The health data is linked to the price data via the location of the healthcare facilities visited by the young people. According to the authors of the study, this generally also represents the place of residence of the adolescents, as patients are assigned to healthcare facilities based on their home address. The allocation also appears to be reliable in that only adolescents who visited a doctor in the same city during the six-year study period are taken into account. However, individual cases, such as regular check-ups at specialist clinics in other cities, cannot be ruled out.

Socioeconomic factors of the cities are also recorded, including the proportion of households in extreme poverty and the proportion of the population in informal employment. A complete overview of all data sources can be found in the Appendix of the study.

Model Specification

The study uses multivariate regression models (see Appendix A.2.2) to investigate the effects of sugar-sweetened beverage taxation on BMI percentiles[72] as well as on the risk and prevalence of overweight and obesity. Multivariate regression models enable the simultaneous analysis of several target variables. It should be noted, however, that this procedure does not allow causal statements unless the underlying relationships have already been proven to be causal.

The model assumes a linear relationship between price changes and body measurements without empirical evidence, which greatly simplifies reality. Logarithmized beverage prices with a lag of one and two years are used as predictors to capture delayed responses. To control for general trends in the results and country-wide influences, fixed effects are included for the observed years. In addition, the model integrates control variables at the individual, clinical and city level, including age, frequency of physician visits and the diagnosed health conditions. Despite the consideration of numerous variables, potentially relevant factors that could influence the relationship between price changes and body measurements are not taken into account. These include comprehensive substitution effects (e.g., price development of other calorie-containing beverages), environmental conditions (e.g., access to green spaces), physical activity or initiatives (e.g., awareness campaigns, measures

[71] including three cities with price reductions after the introduction of the tax
[72] based on standardized BMI z-scores

for healthier eating or more physical activity in schools). In addition, the heterogeneity within the sample is neglected, as no individual differences, for example with regard to socioeconomic status or the base-level consumption of sugar-sweetened beverages, are taken into account.

Communication of Results

The presentation of results is generally complete and includes common statistical parameters such as confidence intervals and p-values in the continuous text. However, no p-values are reported in the results tables, which makes it difficult to classify the results. Changes in the weight-related variables, such as the prevalence of overweight and obesity, are reported both in absolute terms (in relation to the total sample) and in relative terms (in relation to the respective group). Where necessary, the values are contextualized (e.g., conversion of BMI reduction into weight loss), which helps to interpret the results.

The methodology of the study as well as missing aspects are explained in detail in the online appendix of the study. The validity of the results is supported by the comparison with previous studies, in which the reported findings include consistently. To check the robustness of the main results, several sensitivity analyses (see Appendix A.3.2) are performed. These include adjustments for regionally specific linear time trends, consideration of lagged responses to possible substitutes, alternative body weight targets, and sample adjustments. Although the results prove to be robust to these varied assumptions, possible model-related uncertainty in the results due to other key methodological assumptions (e.g., relevant influencing factors not taken into account, assumption of a linear relationship between price changes and body measurements) cannot be ruled out.

The study authors state that they follow reporting guidelines for observational studies. They also transparently name several limitations of their study. These include the limited generalizability of the results to the entire Mexican youth population due to the lack of representativeness of the sample. However, this limitation should be extended even further: The transferability of these results to other countries is also questionable, as the prevalence of overweight and obesity among adolescents in Mexico before the tax was 46.1 %, while in Germany during the same time period, for example, it was only 15.4 % among children and adolescents (Bundesministerium für Gesundheit, 2024). Furthermore, the study authors report on the limitation of not taking into account the heterogeneity of the sample due to a lack of available data, the use of only approximate prices and the incomplete consideration of factors that influence the level of price changes. The authors of the study openly point out that, despite extensive sensitivity analyses, no causal statements can be derived from the results. They consistently and correctly speak of correlations and not cause-and-effect relationships. Overall however, the narrative structure sometimes gives the impression of greater certainty in the results than is actually the case. This is reflected, for example, in the relatively generalized conclusion that large price increases could be accompanied by observable changes in weight-related parameters. However, the simulated results of this underlying statement are based exclusively on specific subgroups of the sample (e.g., girls or overweight girls) and on certain conditions (e.g., when prices are lagged by two years, but not after one year).

Summary of the Limitations of the Study by Gračner et al. (2022)

Data Foundation

- **Data Sample:** Due to the non-representative sample of Mexican adolescents (only those living in urban areas and insured by a specific company, thus presumably socioeconomically advantaged; additionally more likely to be younger, heavier, and female), the study results are biased and therefore only limitedly generalizable to the entire adolescent population of Mexico.
- **Other Input Data:** Due to the lack of price data availability, sales are considered only carbonated sugary drinks not within the service industry outlet, which may lead to potential discrepancies from the actual overall price change.

Model Specification

- **Model Methodology:**

  Causality and linear relationships are assumed in multivariate regression analyses beforehand during model construction, but not proven by the model itself.
- **Influencing and Target Variables:** The assumption of a linear relationship between price changes and body measurements without supporting evidence creates a highly simplified representation of real phenomena and their relationships.
- **Other Influencing Factors:** Due to inadequate consideration of substitution effects, neglect of sample heterogeneity (e.g., socioeconomic status, baseline consumption), and failure to account for environmental conditions, physical activity, or initiatives, the simulated effects of the tax may not necessarily exist in their presented form.

Communication of Results

- **Presentation of Results:** The partially incomplete presentation of p-values complicates the interpretation of some results.
- **Model Evaluation:** Despite validation of the results with previous studies and support through various sensitivity analyses, there is ongoing uncertainty in the results due to the failure to account for various potentially relevant influencing factors.
- **Contextualization of Study Results:** The authors' conclusion (large price increases being associated with observable changes in body weight parameters) gives the impression of high certainty, although the evidence in the modeling study was found only for certain subgroups (girls) and under specific conditions (after two years).

Overall Assessment of the study by Gračner et al. (2022)

The study by Gračner et al. (2022) is based on data from a non-representative sample of Mexican adolescents, which means that the results can only be applied to the entire Mexican adolescent population to a limited extent. A generalization of the results to Germany is not possible due to fundamental differences in the body measurements of young people, among other things. Furthermore, the model used in the study only depicts the complex reality in a simplified form, as it ignores significant influencing factors such as the heterogeneity of the sample or extensive substitution effects. Despite validation of the results through previous studies and additional sensitivity analyses, a certain degree of uncertainty remains. Overall, the reliability of the results must be critically questioned in view of the aforementioned limitations, which is why the study should not be used as a reliable basis for political decisions.

4.6 Study by Basto-Abreu et al. (2019), Mexiko

Full reference source:

Basto-Abreu, A., Barrientos-Gutiérrez, T., Vidaña-Pérez, D., Colchero, M., Hernández Fernández, M., Hernández-Ávila, M., Ward, Z. J., Long, M. W., & Gortmaker, S. L. (2019). Cost-effectiveness of the sugar-sweetened beverage excise tax in Mexico. *Health Affairs*, *38*(11), 1824–1831. https://doi.org/10.1377/hlthaff.2018.05469

Executive Summary

The study by Basto-Abreu et al. (2019) models health and health economic effects of a tax on sugar-sweetened beverages introduced in Mexico in 2014, as well as the potential effects of a hypothetical doubling of the tax rate using a cohort simulation. The study authors conclude from the simulation results that the introduction of taxes in both scenarios leads to a reduction in cases of obesity, which could prevent diseases, improve quality of life and save on healthcare costs. However, the simulation results are based on a questionable data basis that does not realistically depict the actual purchasing and consumption behavior of the Mexican population. In addition, highly simplistic assumptions are made, e.g., regarding tax effects. The reliability of the study results must therefore be critically questioned. Furthermore, the model used is not able to provide evidence that the tax is actually causally responsible for these effects. Overall, from a statistical-methodological perspective, the study does not meet the required quality standards to serve as a reliable basis for evidence-based decisions.

4.6.1 Summary of the Study

The content of the study by Basto-Abreu et al. (2019) is summarized below.

Research Focus

The study simulates the effects of the specific consumption tax on sugar-sweetened beverages introduced in Mexico in 2014, for children and adults over a period of ten years. The potential impacts of the tax on health-related and health-economic aspects are simulated in two scenarios: (I) the actual scenario of the introduced specific consumption tax of 1 Mex$ per liter of sugar-sweetened beverages and (II) a hypothetical scenario of a specific consumption tax of 2 Mex$ per liter.

Methodology

The estimation of the tax impact is based on demographic data, as well as nutrition and health-related characteristics of the Mexican population, population projections, data on obesity-related diseases and their mortality, and healthcare expenditure data. Using a cohort simulation model, changes in the BMI distribution of the Mexican population in 2014 (the time of the tax introduction) from ages 2 to 100 are forecasted over a period of ten years – considering both the assumption of constant consumption of sugar-sweetened beverages and under the assumption of consumption changes due to a tax of 1 and 2 Mex$ per liter. The assumed consumption changes are converted into changes in BMI using a change-in-change analysis (see Appendix A.3.4). The forecasted BMI distributions estimate the cases and rates of obesity, avoided disease cases (e.g., diabetes), quality of life indices, and savings in healthcare costs. Sensitivity analyses examine the effects of an alternative discount rate of 5 % instead of the 3 % used in the main analysis on healthcare costs.

Results

The study forecasts the following changes due to the 1 Mex$ per liter tax on sugar-sweetened beverages over ten years: (1) a reduction in the prevalence of obesity by 239,900 cases, 95 % CI [173,000; 306,000], 94,300 of which are children, 95 % CI [51,600; 137,900], (2) a reduction in the incidence of various diseases, including a reduction of about 61,340 cases of diabetes, 95 % CI [31,940, 95,170], (3) an improvement in quality of life, such as 918 gained life years, 95 % CI [493; 1,420], and (4) healthcare cost savings of about 92 million US$[73], 95 % CI [47; 148], which would correspond to a saving of nearly 4 US dollars per US dollar spent on the tax, 95 % CI [2; 6]. For the hypothetical scenario of doubling the tax to 2 Mex$ per liter of sugar-sweetened beverages, the study forecasts approximately a doubling of these effects.

Conclusions of the Study Authors

The study authors conclude from their forecasts that the tax on sugar-sweetened beverages in Mexico is an effective measure to reduce cases of obesity and related diseases in both children and adults, which leads to an improvement in quality of life and savings in healthcare costs. Additionally, they interpret that increasing the tax in Mexico could further strengthen the positive effects. They speculate that countries with similar conditions could benefit from implementing a comparable tax.

4.6.2 Evaluation of the Study

The study by Basto-Abreu et al. (2019) is evaluated below based on the assessment criteria presented in Chapter 3.

Data Foundation

The analysis relies on a synthetic cohort developed based on national statistics, representing the Mexican population of 2014 from ages 2 to 100. The development of this cohort in terms of health-related factors is simulated based on official population projections over a period of ten years, with the prognosis ending at death or reaching the age of 100. Data from various external studies are used, which are evaluated below.

The baseline consumption data for sugar-sweetened beverages come from a national health and nutrition study (ENSANUT; Instituto Nacional de Salud Pública, 2012), where participants were asked to report the amount of beverages consumed in the last 24 hours using a semi-quantitative questionnaire. While the data accounts for age and gender-specific differences, their overall validity is questionable, as the survey is limited to a short time period. Additionally, self-reported data can be subject to inaccuracies, such as recall bias, estimations, or the desire to provide socially acceptable responses.

The data on the change in purchasing behavior regarding sugar-sweetened beverages as a result of the tax stem from the study by Colchero, Rivera-Dommarco et al. (2017). This study compares the actual purchasing data from the two years following the tax introduction (2014 and 2015) with a counterfactual scenario without the tax, based on trend forecasts from real data from 2012 to 2013. The data reflects the total purchase volume

[73] Costs were reported in US$, not in Mex$.

of a household but does not consider individual differences. Consequently, the change in purchasing behavior regarding sugar-sweetened beverages (price elasticity of demand) is derived as a flat (percentage) value across all population groups. The reliability of the results is limited, as the heterogeneity in terms of age and gender is completely neglected. Additionally, the beverage purchases were calculated based solely on retail data, without considering data from the gastronomy sector, which leads to potential bias in the results. It is also questionable to what extent the data, which is solely based on urban areas, is representative of the entire population, as there is evidence that the decline in consumption of sugar-sweetened beverages in rural areas following the tax introduction was less pronounced (Colchero, Molina & Guerrero-López, 2017).

The conversion of changes in consumption into weight or BMI changes is done using energy balance equations through a Change-in-Change analysis. According to the study authors, the conversion factors used account for other factors such as physical activity or caloric compensation. The conversion factors for adults are based on a study of weight loss in Mexican female teachers. Since a single professional and gender group is not representative of the general population, biases are to be expected when extrapolating the results to the entire adult population. The conversion factors for children and adolescents are based on a study with children in the Netherlands, whose applicability to Mexico is questionable. Moreover, the Dutch study only examined children aged 4 to 11 years (de Ruyter et al., 2012), while Basto-Abreu et al. (2019) apply the results to children and adolescents aged 2 to 19 years.

Both the purchase data and the conversion factors for changes in weight based on consumption used by Basto-Abreu et al. (2019) are non-experimental (as also criticized by Thiboonboon et al. (2024)[74]), even though data from randomized controlled trials – assuming their applicability – are known to be more reliable and lead to less uncertainty in the estimates. The limitations of non-experimental studies lie primarily in their limited causal validity, as effects are observed only subsequently (Doering & Bortz, 2016).

The study forecasts the impacts of BMI changes on health-related aspects based on relative risks from epidemiological studies. Extrapolation is used in the absence of data for certain age groups, which can lead to biases as assumptions regarding disease rates and mortality may not reflect actual trends. Moreover, it is assumed that disease and mortality rates remain constant over the entire forecast period, neglecting potential medical advancements as well as epidemics or pandemics.

Treatment costs are based on national data from Mexico, particularly from health ministries. If such data is not available, U.S. treatment costs are used, adjusted for the differences in medical service costs between the two countries. Despite calibration, U.S. data may not fully reflect the specifics of the Mexican healthcare system.

Model Specification

The study uses a cohort simulation for the modeling method (see Appendix A.1.1). Its main limitation (as with most non-experimental studies) is that causal relationships cannot be proven. Therefore, health-related and/or health-economic effects calculated in the study cannot be directly attributed to the introduction of the tax (even if this may seem plausible), as effects could also be due to concurrent changes or developments.

[74] The study by Basto-Abreu et al. (2019) is one of the 14 studies reviewed in the systematic review by Thiboonboon et al. (2024), which identifies methodological challenges in studying the economic impacts of taxes on sugar-sweetened beverages.

The study operationalizes changes in the consumption of sugar-sweetened beverages by changes in purchasing behavior for these beverages, but purchasing behavior cannot be unconditionally equated with actual consumption, as the purchased beverages do not necessarily have to be fully consumed. The use of data on changes in purchasing behavior[75] also implies the assumption that the tax effect fully takes place two years after its introduction and remains constant over the entire ten-year forecast period, even though sufficient empirical evidence for this is lacking. This assumption neglects potential short-term fluctuations, which would only reflect initial reactions to the tax and not the long-term effectiveness (e.g., consumer behavior adjustments). Furthermore, factors such as seasonal fluctuations or time-limited educational campaigns, which may have influenced consumption during these two years, are not considered in the calculations.

The forecast of the cohort's development in terms of BMI and health-related factors is performed using probabilistic sensitivity analyses (see Appendix A.3.2).[76] The calculations were based on predetermined distributions for the model's key parameters. Details on the selection and distribution of the parameters are not documented. It also remains unclear how variations in the parameters affect the results. A more detailed presentation of parameter variation and iteration results would contribute to the transparency of the modeling and the assessment of uncertainties.

The model specification does not consider all relevant influencing factors. For example, Basto-Abreu et al. (2019) ignore substitution effects: in addition to the decline in the consumption of taxed beverages, an increase in the consumption of non-taxed beverages, which may also contain sugar, was observed in Mexico (Aguilar et al., 2021; Colchero, Rivera-Dommarco et al., 2017).[77] This suggests that taxed beverages were partially substituted by other calorie-containing beverages, so the overall caloric intake from beverages may have decreased less than simulated by the study authors. Therefore, the effect of the tax could be overestimated. Furthermore, interactions with parallel reforms, such as the taxation of calorie-dense solid foods in Mexico, are not considered. A study investigating the effects of both taxes (on beverages and foods) in Mexico showed that there was significant substitution between taxed and non-taxed categories, and that no statistically significant reduction in the total calories purchased could be observed (Aguilar et al., 2021), which fundamentally challenges both the assumptions made by Basto-Abreu et al. (2019) and likewise the resulting outcomes. Inflation, which significantly affects long-term costs and economic efficiency, also appears to be neglected. Socioeconomic status is similarly not included. The consideration of these factors would have allowed for the analysis of the tax's effects on different population groups, and to account for differences in price sensitivity, access to healthcare, and health awareness.

Communication of Results

The estimated decrease in obesity prevalence and health-related effects are consistently expressed in absolute numbers of cases. In the case of obesity prevalence, the effects are also expressed in percentage points, which allows for a better interpretation. However, the prevented illness cases or improved life years are not presented as so. Data in natural or relative frequencies would have increased comprehensibility. There is also no contextualization

[75] Colchero, Rivera-Dommarco et al. (2017) simulate a decrease in the purchase of sugar-sweetened beverages by 5.5% in 2014 and 9.7% in 2015. Basto-Abreu et al. (2019) infer a constant average decline of 7.6% over the entire ten-year study period based on this.

[76] 10,000 iterations are conducted to forecast changes in BMI, and 100,000 iterations are performed to estimate health-related changes.

[77] e.g., 100% fruit juices

of the saved health costs given for the ten-year study period. All calculations are supplemented with confidence or uncertainty intervals[78] which transparently illustrates the variability.

The validation of the estimates for the reduction in obesity prevalence in adults shows discrepancies when compared to other studies. The study authors attribute these deviations to the fact that other approaches did not take into account caloric compensation. The assumptions made by Basto-Abreu et al. (2019) are therefore considered more realistic than those in other approaches. However, the estimated reduction in diabetes prevalence aligns with results from other studies. Differences arise in the estimated cost savings: Basto-Abreu et al. (2019) state that their model produces conservative results and that the discrepancies with other studies' results can primarily be explained by different assumptions regarding baseline costs. Since the differences are significant, it remains questionable whether and to what extent costs are actually saved, especially due to the questionable data foundation.

To forecast potential savings in healthcare costs, a sensitivity analysis is conducted to assess the impact of the discount rate[79] on the results. While conducting such sensitivity analyses is generally useful, the models in Basto-Abreu et al. (2019) are based on potentially distorted sales data for sugar-sweetened beverages. Therefore, it is doubtful whether such a sensitivity analysis as the final model step – and the results in general – would provide meaningful results.

The study transparently identifies some methodological limitations, such as the questionable data quality regarding baseline consumption, which, according to the study authors, tends to lead to conservative estimates of the tax effect. Additional problems are highlighted such as with data availability for changes in consumption and the simplified assumption of a uniform rather than heterogeneous BMI change. However, the role of possible other influencing factors (e.g., substitution, parallel political measures) that could explain part of the effect or put the results into perspective is not adequately discussed.

The study authors conclude from the simulation results that the introduction of the tax would lead to a reduction in obesity cases, preventing diseases, improving quality of life, and saving healthcare costs. However, this conclusion is overly generalized for two reasons: (1) The significant study limitations, particularly regarding the data foundation and simplifying model assumptions, reduce the reliability of the statement and should be transparently communicated. (2) Although the relationships seem plausible, the model used is not capable of causally proving that the observed effects are definitively due to the tax.

Furthermore, the study authors argue that countries with similar baseline conditions could also benefit from introducing a tax. However, since country-specific differences for the baseline conditions can be diverse and complex (e.g., demographic, social, and economic conditions, consumption behavior, price elasticities, and healthcare systems), all prerequisites must be carefully and critically assessed before transferring the results. Applying this to Germany seems highly implausible, given the differing deviations regarding BMI in children, as discussed in Section 4.5.2.

[78] The uncertainty interval is a range of possible values of a variable that takes into account all types of uncertainties, including random and systematic errors and subjective judgments.

[79] An alternative discount rate of 5 % is used instead of the 3 % rate used in the main analysis.

Summary of the Limitations of the Study by Basto-Abreu et al. (2019)

Data Foundation

- **Data Sample:** Due to the use of national population data and official population projections, there are no limitations regarding the representativeness of the sample.
- **Other Input Data:** Data does not adequately reflect consumption and purchasing behavior (baseline consumption based on self-reports and a short survey period; urban beverage purchasing data without considering the gastronomy sector as well as age and gender-specific differences) and thus biases results; non-experimentally collected conversion factors from consumption to weight changes based on specific populations (adults: female teachers; children: Dutch data), making the transferability to the general population questionable due to physiological or ethnic differences; compensating for missing data (health-related aspects: extrapolation; treatment costs: US data) does not adequately reflect the Mexican healthcare system and may skew results.

Model Specification

- **Model Methodology:** Causality is assumed beforehand in cohort simulations during model construction, but not proven by the model itself; distribution parameters for probabilistic sensitivity analyses are not disclosed, which limits transparency.
- **Influencing and Target Variables:** Simplified assumptions (constant tax effect after two years over 10 years neglecting fluctuations and adjustments, equating consumption with purchasing behavior) do not accurately represent real phenomena and their relationships.
- **Other Influencing Factors:** Neglecting substitution effects and interactions with parallel reforms (even though evidence suggests no change in the total amount of calories purchased), makes the reliability of the results questionable; failure to consider other relevant factors (e.g., seasonal fluctuations, educational campaigns, inflation, socioeconomic status), means that the simulated effects may not necessarily exist in the presented form.

Communication of Results

- **Presentation of Results:** Lack of contextualization of results complicates their interpretation.
- **Model Evaluation:** Discrepancies in the results compared to other studies (obesity prevalence, cost savings) due to differing assumptions, with actual values remaining questionable; due to distorted consumption and purchasing data, the relevance of sensitivity analyses in the final modeling step (and for the results in general) is limited.
- **Contextualization of Study Results:** Inadequate discussion of unconsidered influencing factors; conclusions (tax leads to a reduction in obesity, diseases, and healthcare costs; transferability to countries with similar conditions) are overly generalized due to the failure to consider study limitations.

Overall Assessment of the Study by Basto-Abreu et al. (2019)

The reliability of the study results by Basto-Abreu et al. (2019) is particularly questionable due to the insufficient data foundation, oversimplified model assumptions, and the failure to consider relevant influencing factors. The basis for the data shows significant shortcomings: The data on baseline consumption appears to be error-prone due to the data collection method and the research question. The purchasing data for sugar-sweetened beverages excludes the gastronomy sector and neglects age and gender-specific differences – therefore, it is skewed and generalized and does not adequately reflect real purchasing behavior. Additionally, the conversion from consumption to weight changes is based on specific populations (adults: female teachers; children: Dutch data), which cannot be generalized to the entire Mexican population due to physiological and ethnic differences. Health-related effects are simulated based on this fragile foundation, which excludes the possibility of drawing reliable conclusions. The highly simplified assumptions (e.g., full tax effect after two years, constant tax effect over 10 years, equating consumption and purchasing behavior) and the neglect of substitution effects and parallel political measures further undermine the validity of the results and conclusions. Overall, the study is by no means a reliable basis for political decisions.

5 Summary and Discussion

In this expert opinion, six different modeling studies regarding the effects of a tax on sugar-sweetened beverages have been evaluated from a statistical-methodological perspective. This chapter summarizes the key limitations of the individual evaluations, discusses them, and derives key takeaways from them.

5.1 Summary of the Study Evaluation

The key limitations regarding the data foundation, model specification, and the communication of results of the six evaluated studies are summarized below. For the sake of readability, only the names of the study authors are provided, without listing the publication year of each respective study. In particular – unless explicitly marked otherwise – the author reference „Emmert-Fees et al.“ refers to the study by Emmert-Fees et al. (2023), which was evaluated as one of the six studies in the report, and not to Emmert-Fees et al. (2024).

Assessment of the Data Foundation

- **Data sample:** The two evaluated studies based on real samples structurally deviate from the target population of interest: In the study by Rogers, Cummins et al., overweight and obese girls are underrepresented, putting into question the simulated effect for sixth-grade girls; Gračner et al., on the other hand, only examine adolescents living in urban areas and insured by a specific company. This group differs in age, weight, and gender distribution from the average Mexican youth population, and is also likely socioeconomically privileged. Consequently, the results are only partially transferable to the entire adolescent population of Mexico. **Since the samples in these two studies do not adequately represent the population of interest, their results should be considered unreliable.** In contrast, the evaluated studies based on synthetic samples derived from current official population statistics (Emmert-Fees et al.; Cobiac et al.; Basto-Abreu et al.; Schwendicke and Stolpe) can generally be assumed to match the target population structure. It should be noted however, that the analysis by Schwendicke and Stolpe from 2017 is based on population data from 2012, which may differ from the current German population due to significant demographic changes such as migration since then.

  Aside from the study by Basto-Abreu et al., which models a cohort aged two to 100 years, and thus covers all age groups relevant to the research context, the other studies focus on specific age groups (Emmert-Fees et al.: 30 to 90 years; Schwendicke and Stolpe: 15 to 79 years; Rogers, Cummins et al.: preschool children and sixth-graders; Cobiac et al.: 0 to 17 years; Gračner et al.: 10 to 18 years). Therefore, the conclusions are valid only within the examined age groups and should not be generalized beyond that. Nevertheless Rogers, Cummins et al. – although their sample covers only specific age groups of children – draw conclusions about children or older children in general.

- **Other Input Data:** Often, the input data is not or only partially **transferable** to the study's application context. This is partly because **data from other countries**, due to country-specific differences, does not fit the study sample (Emmert-Fees et al.: prevalence of coronary heart disease and stroke from the UK; Schwendicke and Stolpe: price elasticities from the USA; Basto-Abreu et al.: conversion factors for consumption to weight changes in children from the Netherlands). In some cases, the data is based on **specific subpopulations** and therefore cannot be generalized to the entire population. For instance, Basto-Abreu et al. apply data from female teachers to convert consumption to weight changes to adults in general. Moreover, data is often not available in sufficient **detail** to adequately reflect the heterogeneity of the population. For example, Basto-Abreu et al. assume generalized consumption changes across all age and gender groups; Rogers, Cummins et al. and Cobiac et al. can only make rough regional classifications of socioeconomic status. Furthermore, the data does not always **accurately reflect reality**. For example, data on the purchase of sugar-sweetened beverages in several of the evaluated studies is limited to stationary retail through the exclusion of the gastronomy sector (e.g., in Cobiac et al., Gračner et al., and Basto-Abreu et al.), which skews the resulting effects.

Additionally, the data is sometimes **outdated**. For instance, the consumption data in Schwendicke and Stolpe and partially in Emmert-Fees et al. come from the NVS II, collected between 2005 and 2007. The data on disease costs in Emmert-Fees et al. even goes as far back as to 1999. Furthermore, some of the data is based on **small samples**, which limits its reliability. For example, the estimation of productivity losses due to strokes in Emmert-Fees et al. is only based on 151 patients. Additionally, due to limited availability, most of the data is based on **non-experimental** data collection methods, which have a lower evidence base than experimental studies. This applies to, for example, the conversion factors between consumption and weight changes in Emmert-Fees et al. and Basto-Abreu et al. If data from **meta-studies** is used (e.g., health data in Cobiac et al.), the quality of the underlying primary data is often unclear, especially regarding possible biases and their fit to the study context. Data based on **self-reports** (e.g., Schwendicke and Stolpe: body indices; Cobiac et al.: beverage purchases and health data; Basto-Abreu et al.: baseline consumption) is error-prone, which can lead to uncertainties in the results. Furthermore, the data is sometimes **incomplete**. For instance, Emmert-Fees et al. lack risk factors for coronary heart disease and strokes. Methods to address missing data (Emmert-Fees et al.: estimation; Basto-Abreu et al.: extrapolation and supplementation with data from other studies) do not always ensure validity. Finally, some datasets are **incompatible** (e.g., consumption data in Emmert-Fees et al.; consumption and body data in Schwendicke and Stolpe), which can lead to discrepancies in results, such as for specific age groups.

The datasets used to derive model parameters and assumptions about the model structure show weaknesses in all evaluated studies – some significantly so. For example, Emmert-Fees et al. use data from a total of 36 different sources for the model parameterization, which contributes to a substantial degree of uncertainty within the model. This suggests that the parameters in all the evaluated studies (to varying degrees) may be biased. **Modeling based on such a fragile foundation does not allow for reliable conclusions.** The poor availability and quality of the data thus have serious implications for the reliability of the results. Emmert-Fees et al. (2024) demonstrated in a methodological follow-up study to the study evaluated in this report by Emmert-Fees et al. (2023), that results from modeling the effects of a tax on sugar-sweetened beverages can vary greatly depending on the data used.

Assessment of the Modeling Specifications

- **Model Methodology:** The modeling methods used (microsimulation in Emmert-Fees et al.; Monte Carlo simulation in Schwendicke and Stolpe; cohort simulation in Cobiac et al. and Basto-Abreu et al.; time series analysis in Rogers, Cummins et al. and Cobiac et al.; multivariate regression analysis in Gračner et al.) are generally suitable for comparing real situations with hypothetical scenarios. However, causality is assumed beforehand in all methods during the model construction, but it is not proven by the model itself. **If a causal relationship has not already been proven in advance – an aspect that is not explicitly addressed in any of the studies – the modeling procedure does not provide a basis for making the tax *causally* responsible for the simulated effects.** Similarly, the assumption of linearity is assumed for regression models, but is not demonstrated by the model.

In time series analyses, there is also the risk of biased effect estimates (which cannot be completely ruled out), as the calculation of the simulated trend is sensitive to the time of intervention. In addition, 100 simulation runs per group in a Monte Carlo simulation, as carried out by Schwendicke and Stolpe, are not sufficient for reliable results. The replicability of the modeling is also made more difficult in some studies by insufficient information in the model specification. Examples include the unclear selection and distribution of parameters in Schwendicke and Stolpe and Basto-Abreu et al. as well as missing information on ARIMA properties and unchecked assumptions in Rogers, Cummins et al.

- **Influencing and Target Variables:** The influencing variables are often inadequately operationalized. For example, purchase is frequently equated with consumption (Cobiac et al.; Basto-Abreu et al.). In addition, the heterogeneity of consumption behavior regarding age and gender-specific differences is often not taken into consideration (Emmert-Fees et al.; Schwendicke and Stolpe; Cobiac et al.; Basto-Abreu et al.). Both aspects lead to a strong generalization of the results.

  The relationships between influencing and target variables are often modeled in a simplified – mostly linear – way and without sufficient empirical evidence. One example is the assumption of a direct correlation between changes in consumption and weight (Schwendicke and Stolpe; Cobiac et al.), which largely ignores physiological differences, physical activity and caloric compensation and thus contradicts the multifactorial development of overweight and obesity. All studies that model health-related effects (Emmert-Fees et al.; Schwendicke and Stolpe; Cobiac et al.; Basto-Abreu et al.) make assumptions about the temporal effectiveness of the tax, but fail to take into account both possible fluctuations and long-term adaptation effects in consumer behavior. The other two studies also make strongly simplifying assumptions: Gračner et al. assume a linear relationship between price changes and body measurements without empirical evidence; Rogers, Cummins et al. assume a linearly continued trend in obesity prevalence. **The simplified assumptions made by all evaluated studies suggest that the model results do not reflect the real phenomena and their relationships with sufficient precision**.

- **Other influencing factors:** Considering all relevant influencing factors presents a significant challenge due to the complexity of the real situation being modeled for the introduction of a tax. **However, the evaluated studies do not adequately address this problem**. Substitution effects and socioeconomic status are either not sufficiently considered, or not considered at all. Parallel measures, such as awareness campaigns or political reforms, are also consistently neglected. One example of this is the study by Basto-Abreu et al., which examines the effects of a tax on sugar-sweetened beverages on body weight, health and health-economic aspects in Mexico. The identified effects are based on the assumption of a reduction in the consumption of sugar-sweetened beverages due to the tax. However, the reliability of these results is questionable, as other studies in Mexico show that when considering other foods, no significant changes in the total calorie purchases is found. **The neglect of important influencing factors carries the risk that the simulated effects do not accurately represent the real relationships, leading to overly generalized or inaccurate conclusions about the effects of the tax.**

Assessment of the Communication of Results

- **Presentation of Results:** One study (Schwendicke and Stolpe) completely excludes confidence intervals. In the other studies, **the confidence intervals inadequately capture the variability of the simulated results**, because they only reflect random errors but do not account for systematic errors. The absence of *p*-values, effect sizes (Rogers, Cummins et al.; Cobiac et al.; Gračner et al.), and trend parameters (Rogers, Cummins et al.) further complicates a robust assessment of the statistical significance of the results. Moreover, some studies lack proper contextualization, especially when presenting effect sizes such as reductions in disease prevalence or cost savings. For instance, Emmert-Fees et al. estimate a health cost reduction of approximately three billion euros in three scenarios. However, this corresponds to a relative cost reduction of only under 0.04 % – an information that is not communicated by the authors and should be considered small in the context of total costs. **Solely presenting absolute values, suggests to the reader a stronger effect than actually present.** This issue similarly occurs in the studies by Schwendicke and Stolpe, Rogers, Cummins et al., Cobiac et al., and Basto-Abreu et al. Information presented in natural or relative frequencies would have noticeably improved the comprehensibility of the results.

- **Model Evaluation:** Systematic analyses of (global) uncertainties in the modeling results are largely absent and typically limited to sensitivity analyses. These analyses in all studies only consider variations in individual parameters without varying central model assumptions such as significant influencing factors or the assumption of linear relationships. Therefore, they are only partially meaningful and, due to the insufficient data basis, inherently questionable. Cobiac et al. acknowledge the sensitivity of the results to minimal changes in input data but do not simulate their impact. In Emmert-Fees et al., a sensitivity analysis (tiered tax, 10 % sugar reduction in sugar-sweetened beverages) results in simulated effects smaller than in the three main scenarios, thereby relativizing the results of the main analysis.

  In Cobiac et al., validity analyses show discrepancies in sugar consumption and the reduction in sugar consumption, indicating an overestimation of the effects. Similar deviations appear in the obesity prevalence and cost savings in Gračner et al., with the actual values remaining questionable. **Uncertainties and possible biases in the reported results are generally insufficiently addressed in all studies.**

- **Contextualization of Study Results: The interpretation of model results often occurs without comprehensive consideration of the study limitations and uncertainties in the results, meaning the reliability of the results in all studies must be critically questioned.** Although study limitations are usually discussed broadly, they are often only superficially addressed and rarely reflected upon in terms of their potential impact on the results. **This leads to overly generalized conclusions.** For example, Emmert-Fees et al. present the tiered tax as the optimal tax form but do not consider the uncertainties highlighted in a sensitivity analysis. Rogers, Cummins et al. and Gračner et al. report effects only for specific subgroups, such as girls, but generalize these effects to boys as well. Cobiac et al. interpret a slight increase in life expectancy – ranging from a few days to at most a month – as a significant improvement, overinterpreting the actual relevance of the effects. Particularly problematic are incorrect causal conclusions: Schwendicke and Stolpe, Cobiac et al., and Basto-Abreu et al. conclude that a tax would reduce obesity, although this is not supported by the models used. Such misleading conclusions are especially critical, as they are often picked up by the media and used to argue for regulatory measures.

5.2 Key Findings

On the basis of existing research on modeling studies and the evaluations of the individual studies carried out in this report, the following key findings can be derived:

1. Modeling studies that investigate (hypothetical) regulatory measures, e.g., the introduction of a sugar tax in Germany, are not independent evidence of causal relationships; rather, the existence of causal effects between influencing and target variables is assumed beforehand in the chosen modeling methods.
2. The existence and quantification of causal mechanisms is based on assumptions, e.g., the transferability of earlier empirical studies. Such assumptions are inevitably associated with uncertainties in the modeling results. Every decision for a certain assumption is at the same time a decision against alternative possibilities. Modeling results are therefore significantly influenced by the underlying assumptions.
3. The validity of modeling studies is heavily dependent on the availability and quality of the required empirical input data. Modeling studies usually obtain these from external sources and assume their representativeness and accuracy, often without explicitly checking either of them.
4. Inaccurate, biased and/or non-representative data, as well as incorrect or overly simplistic assumptions, lead to model results that may be consistent within the model, but do not accurately reflect reality.
5. Even in cases where the assumptions are fundamentally correct and the input data is representative, the results of modeling studies are always subject to uncertainties that could be considerable, and are often ignored in public health issues.

5.3 Discussion

Political decisions should be based on an informed discourse regarding goals and measures. Science can contribute by providing evidence on causes and effects, for example, through methodologically appropriate, systematic evaluations. To ensure an informed debate, the data foundation, evaluation methods, and results must be accessible and transparent (Leopoldina, 2021).

The careful consideration of the purpose for which the results of modeling studies can be used (“fit for purpose”) is a key element by which they are assessed. To avoid improper application, the results must be presented and communicated transparently. It is important to clearly distinguish between results that serve the exploration of potential research questions or knowledge gain within a discipline, and those that are to be used for political decisions, such as in legislative processes (Münnich, in press). The former may have lower requirements for the quality of the underlying data, but these should be documented accordingly. However, when study authors claim to provide data foundations for the derivation of political measures, they must meet the strictest criteria regarding data quality, methodological suitability, and adherence to scientific ethical principles.

Given the potentially far-reaching impacts on society and the economy, it is crucial that all involved stakeholders – particularly political decision-makers, financial supporters, and participating researchers – are aware that studies used as the basis for evidence-based policy must meet exceptionally high quality standards at all levels. This applies both to the **data foundation**, which must provide the best possible **representation of real phenomena**, and to the **methods of analysis** used on this data, which must be suitable for adequately modeling the **relationships between these phenomena**. According to current scientific knowledge, modeling studies are fundamentally the best *methods* for investigating the effects of regulatory measures. They can theoretically be suitable for establishing an evidence base for regulation, provided they are based on reliable and carefully validated assumptions and precise data foundations. However, the causality of the modeled mechanisms is not demonstrated by the modeling studies themselves but must be substantiated by corresponding studies beforehand. It must also be possible to estimate the model parameters with sufficient certainty from existing data. The creation of a modeling study suitable for justifying policy measures therefore requires high-quality data relevant to the application context alongside realistic assumptions for modeling the relationships between the data, as well as suitable statistical methods. Only on this basis, results that are convincing in their entirety can be achieved. Inaccurate, skewed and/or unrepresentative data as well as incorrect or oversimplified assumptions about the underlying mechanisms of action lead to model results that may be consistent within the model, but do not correctly reflect reality.

> The report concludes from a statistical-methodological perspective that none of the evaluated modeling studies on the effects of a sugar tax meet the high quality standards required for evidence-based policy decisions.

A core problem is that databases of sufficient quality, which would create a reliable basis for the established models, are not fully available – with this it is not possible to meet an indispensable prerequisite for evidence-based policymaking. The lack of such data is due on the one hand to the complexity of the underlying mechanisms of action, which are subject to influencing factors that are difficult to measure, and on the other hand to unobserved (or even unobservable) heterogeneity of the subpopulations. In the absence of sufficiently detailed data sets, numerous assumptions have to be made. In addition, the sufficiently precise simulation of the population at the micro level is highly challenging. Possible selection errors in data collection, measurement errors due to self-reports or uncertainties due to small samples can have a significant impact on the model results. Study results from other countries are difficult to apply to Germany, as there are always structural differences between the sample of another country and the German population as a whole.

Some of the limitations identified in the report can be barely, or not at all, overcome. This analysis is therefore not intended to diminish the scientific achievements of the authors of the evaluated studies. The criticism is rather directed at the assumption, that such studies can be used without hesitation as a reliable basis for regulatory decisions. This report is therefore intended as an appeal to take the identified limitations adequately into account within the public debate and political decision-making processes. This includes, in particular, clearly quantifying and transparently communicating uncertainties in modeling studies, within the discussion of a possible sugar tax in Germany as well.[80]

[80] A similar conclusion is reached by Emmert-Fees et al. (2024, p. 2) in their method-focused study: „Predicted body weight reductions under SSB taxation are sensitive to assumptions by researchers often needed due to data limitations. Because this variability propagates to estimates of health and economic impacts, the resulting structural uncertainty should be considered when using results in decision-making." The findings of this study provide an indispensable addition to the modeling study by Emmert-Fees et al. (2023) and should therefore be appropriately considered in the debate.

References

Aguilar, A., Gutierrez, E., & Seira, E. (2021). The effectiveness of sin food taxes: Evidence from Mexico. *Journal of Health Economics*, *77*, Article 102455. https://doi.org/10.1016/j.jhealeco.2021.102455

Appelhans, Y. (2024). So sinnvoll wäre eine Zuckersteuer. Retrieved October 7, 2024, from https://www.tagesschau.de/wissen/gesundheit/zuckersteuer-100.html

Arnold, K. F., Harrison, W. J., Heppenstall, A. J., & Gilthorpe, M. S. (2019). DAG-informed regression modelling, agent-based modelling and microsimulation modelling: a critical comparison of methods for causal inference. *International Journal of Epidemiology*, *48*(1), 243–253. https://doi.org/10.1093/ije/dyy260

Athey, S., & Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, *74*(2), 431–497. https://doi.org/10.1111/j.1468-0262.2006.00668.x

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2023). *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung* (17th ed.). Springer.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on nonprobability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Basto-Abreu, A., Barrientos-Gutiérrez, T., Vidaña-Pérez, D., Colchero, M., Hernández Fernández, M., Hernández-Ávila, M., Ward, Z. J., Long, M. W., & Gortmaker, S. L. (2019). Cost-effectiveness of the sugar-sweetened beverage excise tax in Mexico. *Health Affairs*, *38*(11), 1824–1831. https://doi.org/10.1377/hlthaff.2018.05469

Bernabé, E., Vehkalahti, M. M., Sheiham, A., Lundqvist, A., & Suominen, A. L. (2016). The shape of the dose-response relationship between sugars and caries in adults. *Journal of Dental Research*, *95*(2), 167–172. https://doi.org/10.1177/0022034515616572

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Blake, M., Lancsar, E., Peeters, A., & Backholer, K. (2019). Sugar-sweetened beverage price elasticities in a hypothetical convenience store. *Social Science & Medicine*, *225*, 98–107. https://doi.org/10.1016/j.socscimed.2019.02.021

Brown, P., & Jiang, H. (2010). Simulation-based power calculations for large cohort studies. *Biometrical Journal*, *52*(5), 604–615. https://doi.org/10.1002/bimj.200900277

Brown, V., Tan, E. J., Hayes, A. J., Petrou, S., & Moodie, M. L. (2018). Utility values for childhood obesity interventions: A systematic review and meta-analysis of the evidence for use in economic evaluation. *Obesity Reviews*, *19*(7), 905–916. https://doi.org/10.1111/obr.12672

Bundesministerium für Gesundheit. (2024). Förderschwerpunkt Prävention von Übergewicht bei Kindern und Jugendlichen. Retrieved December 12, 2024, from https://www.bundesgesundheitsministerium.de/themen/praevention/kindergesundheit/praevention-von-kinder-uebergewicht.html

Bungartz, H.-J., Zimmer, S., Buchholz, M., & Pflüger, D. (2013). *Modellbildung und Simulation: Eine anwendungsorientierte Einführung*. Springer. https://doi.org/10.1007/978-3-642-37656-6

Cobiac, L. J., Rogers, N. T., Adams, J., Cummins, S., Smith, R., Mytton, O., White, M., & Scarborough, P. (2024). Impact of the UK soft drinks industry levy on health and health inequalities in children and adolescents in England: An interrupted time series analysis and population health modelling study. *PLOS Medicine*, *21*(3), Article e1004371. https://doi.org/10.1371/journal.pmed.1004371

Colchero, M. A., Molina, M., & Guerrero-López, C. M. (2017). After Mexico implemented a tax, purchases of sugar-sweetened beverages decreased and water increased: Difference by place of residence, household composition, and income level. *The Journal of Nutrition*, *147*(8), 1552–1557. https://doi.org/10.3945/jn.117.251892

Colchero, M. A., Rivera-Dommarco, J., Popkin, B. M., & Ng, S. W. (2017). In Mexico, evidence of sustained consumer response two years after implementing a sugar-sweetened beverage tax. *Health Affairs*, *36*(3), 564–571. https://doi.org/10.1377/hlthaff.2016.1231

Cole, T. J., & Lobstein, T. (2012). Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatric Obesity*, *7*(4), 284–294.

Cornelsen, L., Berger, N., Cummins, S., & Smith, R. D. (2019). Socio-economic patterning of expenditures on ‘out-of-home’ food and non-alcoholic beverages by product and place of purchase in Britain. *Social Science & Medicine*, *235*, Article 112361. https://doi.org/10.1016/j.socscimed.2019.112361

de Ruyter, J. C., Olthof,
bibinitperiod M., Seidell, J. C., & Katan, M. B. (2012). A trial of sugar-free or sugar-sweetened beverages and body weight in children. *New England Journal of Medicine*, *397*(15), 1367–1406. https://doi.org/10.1056/NEJMoa1203034

Dickson, A., Gehrsitz, M., & Kemp, J. (2023). Does a spoonful of sugar levy help the calories go down? An analysis of the UK Soft Drinks Industry Levy. *The Review of Economics and Statistics*, 1–29. https://doi.org/10.1162/rest_a_01345

Doering, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5th ed.). Springer.

Emmert-Fees, K. M. F., Amies-Cull, B., Wawro, N., Linseisen, J., Staudigel, M., Peters, A., Cobiac, L. J., O’Flaherty, M., Scarborough, P., Kypridemos, C., & Laxy, M. (2023). Projected health and economic impacts of sugar-sweetened beverage taxation in Germany: A cross-validation modelling study. *PLOS Medicine*, *20*(11), Article e1004311. https://doi.org/10.1371/journal.pmed.1004311

Emmert-Fees, K. M. F., Felea, A., Staudigel, M., Ananthapavan, J., & Laxy, M. (2024). The implications of policy modeling assumptions for the projected impact of sugar-sweetened beverage taxation on body weight and type 2 diabetes in Germany. *BMC Public Health*, *24*(1), Article 2013 (2024). https://doi.org/10.1186/s12889-024-19488-5

Emmert-Fees, K. M. F., Karl, F. M., von Philipsborn, P., Rehfuess, E. A., & Laxy, M. (2021). Simulation modeling for the economic evaluation of population-based dietary policies: A systematic scoping review. *Advances in Nutrition*, *12*(5), 1957–1995. https://doi.org/10.1093/advances/nmab028

Ernst, J., Dräger, S., Schmaus, S., Weymeirsch, J., Alsaloum, A., & Münnich, R. (2023). The influence of migration patterns on regional demographic development in Germany. *Social Sciences*, *12*(5), Article 255. https://doi.org/10.3390/socsci12050255

Ethgen, O., & Standaert, B. (2012). Population- versus cohort-based modelling approaches. *PharmacoEconomics*, *30*(3), 171–181. https://doi.org/10.2165/11593050-000000000-00000

European Union. (2020). *European Statistical System handbook for quality and metadata reports*. Eurostat. https://doi.org/10.2785/666412

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2023). Zeitreihen. In L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot & G. Tutz (Hrsg.), *Statistik: Der Weg zur Datenanalyse* (S. 553–578). Springer. https://doi.org/10.1007/978-3-662-67526-7_14

Faulbaum, F. (2022). Total survey error. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 567–584). Springer.

Fredriksson, A., & Magalhães de Oliveira, G. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journal*, *54*(4), 519–532. https://doi.org/10.1108/RAUSP-05-2019-0112

Glenn, N. D. (2005). *Cohort analysis* (2nd ed.). Sage.

Gračner, T., Marquez-Padilla, F., & Hernandez-Cortes, D. (2022). Changes in weight-related outcomes among adolescents following consumer price increases of taxed sugar-sweetened beverages. *JAMA Pediatrics*, *176*(2), 150–158. https://doi.org/10.1001/jamapediatrics.2021.5044

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hancock, C., & Copley, V. (2016). National Child Measurement Programme: Guidance for data sharing and analysis. https://assets.publishing.service.gov.uk/media/5a81927fe5274a2e87dbe58c/ncmp-analysis-guidance.pdf

Harrison, R. L. (2010). Introduction to Monte Carlo simulation. *AIP conference proceedings*, *1204*, 17–21. https://doi.org/10.1063/1.3295638

Heitmann, A. (2024). Zuckersteuer, Quengelware und Frühstücksrichtlinie. Retrieved August 16, 2024, from https://www.zdf.de/nachrichten/ratgeber/zuckersteuer-fruehstuecksrichtlinie-massnahmen-100.html

Helmholtz Zentrum München. (2024). KORA: Kooperative Gesundheitsforschung in der Region Augsburg. Retrieved December 5, 2024, from https://www.helmholtz-munich.de/en/epi/cohort/kora/kora-studienzentrum

Hummel, E., Wittig, F., Schneider, K., Gebhardt, N., & Hoffmann, I. (2013). The complex interaction of causing and resulting factors of overweight/obesity: Increasing the understanding of the problem and deducing requirements. *Ernähungs Umschau international*, *60*(1), 2–7. https://doi.org/10.4455/eu.2013.002

Icks, A., Claessen, H., Strassburger, K., Waldeyer, R., Chernyak, N., Jülich, F., Rathmann, W., Thorand, B., Meisinger, C., Huth, C., Rückert, I.-M., Schunk, M., Giani, G., & Holle, R. (2013). Patient time costs attributable to healthcare use in diabetes: Results from the population-based KORA survey in Germany. *Diabetic Medicine*, *30*(10), 1245–1249. https://doi.org/10.1111/dme.12263

Instituto Nacional de Salud Pública. (2012). Encuesta Nacional de Salud y Nutrición 2012. Resultados nacionales. https://ensanut.insp.mx/encuestas/ensanut2012/doctos/informes/ENSANUT2012ResultadosNacionales.pdf

Iskandar, R. (2018). A theoretical foundation for state-transition cohort models in health decision analysis. *PLOS ONE*, *13*(2), Article e0205543. https://doi.org/10.1371/journal.pone.0205543

Jahn, B., Friedrich, S., Behnke, J., Engel, J., Garczarek, U., Münnich, R., Pauly, M., Wilhelm, A., Wolkenhauer, O., Zwick, M., Siebert, U., & Friede, T. (2022). On the role of data, statistics and decisions in a pandemic. *Advances in Statistical Analysis*, *106*(3), 349–382. https://doi.org/10.1007/s10182-022-00439-7

Jandoc, R., Burden, A. M., Mamdani, M., Lévesque, L. E., & Cadarette, S. M. (2015). Interrupted time series analysis in drug utilization research is increasing: Systematic review and recommendations. *Journal of Clinical Epidemiology*, *68*(8), 950–956. https://doi.org/10.1016/j.jclinepi.2014.12.018

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall. https://www.webpages.uidaho.edu/~stevel/519/Applied%20Multivariate%20Statistical%20Analysis%20by%20Johnson%20and%20Wichern.pdf

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: Interrupted time series analysis. *BMJ*, *350*, Article h2750. https://doi.org/10.1136/bmj.h2750

Krauss, S., Weber, P., Binder, K., & Bruckmaier, G. (2020). Natürliche Häufigkeiten als numerische Darstellungsart von Anteilen und Unsicherheit – Forschungsdesiderate und einige Antworten. *Journal für Mathematik-Didaktik*, *41*(2), 485–521. https://doi.org/10.1007/s13138-019-00156-w

Law, C., Smith, R., & Cornelsen, L. (2022). Place matters: Out-of-home demand for food and beverages in Great Britain. *Food Policy*, *107*, Article 102215. https://doi.org/10.1016/j.foodpol.2021.102215

Laxy, M., & Emmert-Fees, K. (2025). Evidenz zum gesundheitlichen Effekt von zuckergesüßten Getränken und deren Besteuerung: Warum es sich lohnt, wirklich genau hinzusehen! *Monitor Versorgungsforschung*, *18*(1), 48–55. https://doi.org/10.24945/MVF.01.25.1866-0533.2692

Leopoldina. (2021). Was ist evidenzbasierte Politikgestaltung? Retrieved May 16, 2024, from https://www.leopoldina.org/themen/evidenzbasierte-politikgestaltung/politikgestaltung-22/

Li, J., & O'Donoghue, C. (2012). A survey of dynamic microsimulation models: uses, model structure andmethodology. *International Journal of Microsimulation*, *6*(2), 3–55. https://doi.org/10.34196/ijm.00082

Lohr, S. L. (2021). *Sampling: Design and analysis* (3rd ed.). Chapman; Hall/CRC.

Magoley, N. (2024). Zuckersteuer auf Softdrinks: Was sie bringen würde. Retrieved October 7, 2024, from https://www1.wdr.de/nachrichten/zuckersteuer-fragen-antworten-100.html

Mathes, T., Röding, D., Stegbauer, C., Laxy, M., & Pieper, D. (2024). „Interrupted time series“-Studien zur Bewertung der Kausalität von Interventionseffekten. *Deutsches Ärzteblatt International*, *121*(24), Article 800–4. https://doi.org/10.3238/arztebl.m2024.0150

Max Rubner-Institut. (n. d.). Die Nationale Verzehrsstudie II. Retrieved December 5, 2024, from https://www.mri.bund.de/de/institute/ernaehrungsverhalten/forschungsprojekte/nvsii/

Mertens, E., Genbrugge, E., Ocira, J., & Peñalvo, J. L. (2022). Microsimulation modeling in food policy: A scoping review of methodological aspects. *Advances in Nutrition*, *13*(2), 621–632. https://doi.org/10.1093/advances/nmab129

Mittag, H.-J., & Schüller, K. (2023). *Statistik: Eine interdisziplinäre Einführung mit interaktiven Elementen* (7th ed.). Springer. https://doi.org/10.1007/978-3-662-68224-1

Münnich, R. (n. d.). Evidence-based policies and data quality – What is missing? [in press]. In R. Kirchner, U. Schipper & J. Walter (Hrsg.), *Measuring International Economics*. Springer.

Münnich, R. (2020). Qualität der regionalen Armutsmessung – vom Design zum Modell. In B. Klumpe, J. Schröder & M. Zwick (Hrsg.), *Qualität bei zusammengeführten Daten* (S. 7–25). Springer. https://doi.org/10.1007/978-3-658-31009-7_2

Münnich, R. (2023). Discussion of "Probability vs. nonprobability sampling: From the birth of survey sampling to the present day" by Graham Kalton. *Statistics in Transition new series*, *24*(3), 39–41. https://doi.org/10.59170/stattrans-2023-033

Münnich, R., Schnell, R., Brenzel, H., Diekmann, H., Dräger, S., Emmenegger, J., Höcker, P., Kopp, J., Merkle, H., Neufang, K., Obersneider, M., Reinhold, J., Schaller, J., Schmaus, S., & Stein, P. (2021). A population based regional dynamic microsimulation of Germany: The MikroSim model. *methods, data, analyses*, *15*(2), 241–264. https://doi.org/10.12758/MDA.2021.03

National Health Service. (2011). Adult Dental Health Survey 2009 – Summary report and thematic series. Retrieved October 1, 2024, from https://digital.nhs.uk/data-and-information/publications/statistical/adult-dental-health-survey/adult-dental-health-survey-2009-summary-report-and-thematic-series

National Health Service. (2015). Child Dental Health Survey 2013, England, Wales and Northern Ireland. Retrieved October 1, 2024, from https://digital.nhs.uk/data-and-information/publications/statistical/children-s-dental-health-survey/child-dental-health-survey-2013-england-wales-and-northern-ireland

National Health Service. (2016). Health Survey for England, 2015. Retrieved September 25, 2024, from https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/health-survey-for-england-2015

National Health Service. (2023). National Child Measurement Programme, England, 2022/23 school year. Retrieved September 12, 2024, from https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2022-23-school-year

Nawroth, P. P., & Kumar, V. (2024). Zuckersteuer: Plausibilität und Wertigkeit der Studien. *Monitor Versorgungsforschung*, *05/24*, 52–56. Retrieved October 7, 2024, from https://www.monitor-versorgungsforschung.de/abstract/zuckersteuer-plausibilitaet-und-wertigkeit-der-studien/

Office for Health Improvement and Disparities. (2024). National Child Measurement Programme. Retrieved September 12, 2024, from https://www.gov.uk/government/collections/national-child-measurement-programme

Penfold, R. B., & Zhang, F. (2013). Use of interrupted time series analysis in evaluating health care quality improvements. *Academic Pediatrics*, *13*(6), S38–S44. https://doi.org/10.1016/j.acap.2013.08.002

Public Health England. (2018). National Diet and Nutrition Survey. Results from years 7 and 8 (combined) of the Rolling Programme (2014/2015 to 2015/2016). https://assets.publishing.service.gov.uk/media/5acdf009ed915d32a65db8cc/NDNS_results_years_7_and_8.pdf

Robert Koch-Institut. (2024). KiGGS: Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland. Retrieved December 5, 2024, from https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/Kiggs/kiggs_node.html

Rogers, N. T., Cummins, S., Forde, H., Jones, C. P., Mytton, O., Rutter, H., Sharp, S. J., Theis, D., White, M., & Adams, J. (2023). Associations between trajectories of obesity prevalence in english primary school children and the UK soft drinks industry levy: An interrupted time series analysis of surveillance data. *PLOS Medicine*, *20*(1), Article e1004160. https://doi.org/10.1371/journal.pmed.1004160

Rogers, N. T., Pell, D., Mytton, O. T., Penney, T. L., Briggs, A., Cummins, S., Jones, C., Rayner, M., Rutter, H., Scarborough, P., Sharp, S., Smith, R., White, M., & Adams, J. (2023). Changes in soft drinks purchased by British households associated with the UK soft drinks industry levy: A controlled interrupted time series analysis. *BMJ Open*, *13*(12), Article e077059. https://doi.org/10.1136/bmjopen-2023-077059

Romeike, F., & Stallinger, M. (2021). *Stochastische Szenariosimulation in der Unternehmenspraxis: Risikomodellierung, Fallstudien, Umsetzung in R*. Springer.

Saltelli, A., Ratto, M., Terry, A., Andres, Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). *Global sensitivity analysis: The primer*. Wiley. https://doi.org/10.1002/9780470725184

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: A guide to assessing scientific models*. Wiley. https://doi.org/10.1002/0470870958

Sasse, T., & Metcalfe, S. (2022). Sugar tax. Retrieved August 12, 2024, from https://www.instituteforgovernment.org.uk/explainer/sugar-tax

Schmaus, S. (2023). *Methoden regionalisierter dynamischer Mikrosimulationen*. https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/deliver/index/docId/2049/file/Dissertation_Schmaus.pdf

Schüller, K. (2015). *Statistik und Intuition: Alltagsbeispiele kritisch hinterfragt*. Springer. https://doi.org/10.1007/978-3-662-47848-6

Schwendicke, F., Thomson, W. M., Broadbent, J. M., & Stolpe, M. (2016). Effects of taxing sugar-sweetened beverages on caries and treatment costs. *Journal of Dental Research*, *95*(12), 1327–1332. https://doi.org/10.1177/0022034516660278

Schwendicke, F., & Stolpe, M. (2017). Taxing sugar-sweetened beverages: Impact on overweight and obesity in Germany. *BMC Public Health*, *17*, Article 88. https://doi.org/10.1186/s12889-016-3938-4

Siebert, U., Alagoz, O., Bayoumi, A. M., Jahn, B., Owens, D. K., Cohen, D. J., & Kuntz, K. M. (2012). State-transition modeling: A report of the ISPOR-SMDM modeling good research practices task force-3. *Value in Health*, *15*(6), 812–820. https://doi.org/10.1016/j.jval.2012.06.014

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). English indices of deprivation 2015: Technical report. https://assets.publishing.service.gov.uk/media/5a7f24b240f0b62305b85578/English_Indices_of_Deprivation_2015_-_Technical-Report.pdf

Spielauer, M. (2011). What is social science microsimulation? *Social Science Computer Review*, *29*(1), 9–20. https://doi.org/10.1177/0894439310370085

Statista. (2022). Share of household food and drink expenditure in the United Kingdom (UK) from 2000 to 1st quarter 2020, by at-home and out-of-home consumption. Retrieved September 17, 2024, from https://www.statista.com/statistics/941699/in-home-versus-out-of-home-food-and-drink-spending-united-kingdom-uk/

Statistisches Bundesamt. (n. d. a). Fortschreibung des Bevölkerungsstandes (Bevölkerungsfortschreibung). Retrieved December 5, 2024, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Glossar/fortschreibungdes-bevoelkerungsstandes.html

Statistisches Bundesamt. (n. d. b). Kohorte. Retrieved October 29, 2024, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Geburten/Glossar/kohorte.html

Statistisches Bundesamt. (2024). Gesundheitsausgaben in Deutschland. Retrieved August 13, 2024, from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Gesundheitsausgaben/_inhalt.html

The Royal Swedish Academy of Sciences. (2021). Natural experiments help answer important questions. Retrieved October 23, 2024, from https://www.nobelprize.org/uploads/2021/10/popular-economicsciencesprize2021-3.pdf

Thiboonboon, K., De Abreu Lourenco, R., Cronin, P., Khoo, T., & Goodall, S. (2024). Economic evaluations of obesity-targeted sugar-sweetened beverage (SSB) taxes–A review to identify methodological issues. *Health Policy*, *144*, Article 105076. https://doi.org/10.1016/j.healthpol.2024.105076

United Nations. (1954). *Handbook of statistical organization* (1st ed.). Retrieved May 16, 2024, from https://unstats.un.org/unsd/publication/SeriesF/SeriesF_6E.pdf

Verified Metrics. (n. d.). Cohort modeling. Retrieved August 9, 2024, from https://www.verifiedmetrics.com/blog/cohort-modeling

Wegwarth, O., & Gigerenzer, G. (2011). Nutzen und Risiken richtig verstehen. *Deutsches Ärzteblatt*, *108*(11), A–568. https://www.aerzteblatt.de/archiv/81375/Risikokommunikation-Nutzen-und-Risiken-richtig-verstehen

Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung: Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Springer. https://doi.org/10.1007/978-3-642-35012-2

Werny, L. (2024). Eine Zuckersteuer in Deutschland ist längst überfällig. https://www.geo.de/wissen/gesundheit/eine-zuckersteuer-in-deutschland-ist-laengst-ueberfaellig-34869352.html

Weyer, J., & Roos, M. (2017). Agentenbasierte Modellierung und Simulation. *Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, *26*(3), 11–16. https://doi.org/10.14512/tatup.26.3.11

Weymeirsch, J., Ernst, J., & Münnich, R. (2024). Model recalibration for regional bias reduction in dynamic microsimulations. *Mathematics*, *12*(10), Article 1550. https://doi.org/10.3390/math12101550

Winter, Y., Wolfram, C., Schöffski, O., Dodel, R., & Back, T. (2008). Langzeitkrankheitskosten 4 Jahre nach Schlaganfall oder TIA in Deutschland. *Nervenarzt*, *79*, 918–926. https://doi.org/10.1007/s00115-008-2505-3

World Health Organization. (2022). WHO European regional obesity report 2022. https://iris.who.int/bitstream/handle/10665/353747/9789289057738-eng.pdf

List of Abbreviations

A

ARIMA . . . Autoregressive Integrated Moving Average

B

BMI . . . Body Mass Index

C

CHD . . . Coronary Heart Disease

CI . . . Confidence Interval

I

IMSS . . . Instituto Mexicano del Seguro Social

K

kcal . . . Kilocalories

N

NCMP . . . National Child Measurement Programme

n.d. . . . no date

NVS II . . . Nationale Verzehrsstudie II

Q

QALY . . . Quality-Adjusted Life Year

S

SSBs . . . Sugar Sweetened Beverages

SDIL . . . Soft Drinks Industry Levy

T

T2DM . . . Typ 2 Diabetes Mellitus

W

WHO . . . World Health Organization

Glossary

A

Ad Valorem Excise Tax

An ad valorem excise tax is a value tax and is levied as a percentage of the value of a good.

B

BMI (Body Mass Index)

BMI is a widely used indicator for assessing body weight in relation to height. It is calculated by dividing the body weight in kilograms by the square of the height in meters. The BMI serves as a basis for classifying whether a person is considered underweight, normal weight, overweight, or obese.

C

Causality

Causality describes the relationship between cause and effect, i.e. a change in one variable A results in a change in another variable B. Causality implies a *correlation* of these variables. Conversely, however, a correlation of two variables is not sufficient to conclude causality. For example, it is possible that two variables do not have a direct interdependent influence, but both causally depend on a third variable C, resulting in an observed (apparent) correlation.

Confidence Interval

A confidence interval (CI) corresponds to a random interval that, given the known distribution of the sample, has a predefined probability range for the true but unknown estimate. This means that, for a given confidence level, e.g., 95 % (which is often chosen by default), 95 % of the confidence intervals derived from samples would cover the true value.

Correlation

Two variables, A and B, are said to be correlated if there is a statistical relationship between them. If high values of A are associated with high values of B, this is called a positive correlation. Conversely, a negative correlation exists if high values of A tend to be associated with low values of B. An example of positively correlated variables is height and weight in humans: taller individuals tend to weigh more than shorter ones. However, it is crucial to note that correlation does not imply *causality*. For example, there is a positive correlation between the variables "ice cream sales" and "sunburn occurrence," without a causal relationship between them.

Cross-Price Elasticity

The cross-price elasticity (demand) quantifies the percentage change in the quantity demanded of one good in response to a percentage change in the price of another good. A positive cross-price elasticity occurs when a price increase of good X leads to an increase in the demand for good Y, indicating that consumers substitute one good for the other (e.g., substitution of sugary drinks with fruit juices). Conversely, a negative cross-price elasticity occurs when a price increase of good X is associated with a decrease in the demand for good Y, indicating that consumers complement X with Y – when the price of one good (e.g., barbecue items) increases, consumers buy less of it, which automatically leads to less demand for the complementary good (e.g., charcoal).

K

kcal (Kilocalories)

Kilocalories (kcal) are an (outdated) unit for the amount of energy, that is used to describe food. Informally, the kcal is often simply referred to as a calorie (without kilo).

L

Linear Regression

Linear regression is a statistical method for modeling and predicting the values of a variable Y dependent on one or more variables $X_1, X_2, \ldots$. A linear relationship is assumed between Y and X_i, that is described by the equation

$$Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + \varepsilon$$

The coefficients $b_0, b_1, b_2, \ldots$, are estimated so that the sum of the squared error terms ε, which are the deviations between the observed and predicted values, are minimized over all data points. The variable Y is called the *dependent* or *target* variable, while $X_1, X_2, \ldots$ are called the *independent* or *influencing* variables. If the model contains only one predictor, it is called *simple* linear regression. If multiple predictors are included, it is called *multiple* linear regression. If multiple target variables $Y_1, Y_2, \ldots$ are modeled together – transforming the equation into a system of equations – it is referred to as *multivariate* linear regression.

M

Meta-study/Meta-analysis

A meta-study or meta-analysis is a systematic summary and evaluation of results from primary studies (individual research). The goal is to quantitatively integrate earlier research and present the findings in a consolidated manner based on statistical analysis.

P

Population

The population refers to the set of all entities (e.g., persons) about which a statement is to be made (e.g., the entire population living in Germany). Complete data on the population are rarely available in practice. As a rule, data is collected via a (much smaller) *sample* in order to draw conclusions about the characteristics of the population.

Price Elasticity

Price elasticity (of demand) quantifies the percentage change in the quantity demanded of a good in response to a price change of that good.

Q

QALY (quality-adjusted life year)

The QALY is a measure for assessing a year of life in terms of health and well-being. A year of life in full health or with maximum well-being is assigned a value of 1, while a state without health or well-being, such as death, is assigned a value of 0. For states with limitations, the value lies between 0 and 1, depending on the severity of the impairment. This concept allows to give more weight to life years with higher quality of life, for example, in the context of evaluating medical interventions, than to those of lower quality of life. However, there is no uniform definition for assigning QALY values to specific states of health. Instead, the evaluations are usually based on surveys that capture the subjective well-being of those affected. This data is typically integrated into models to assign an appropriate QALY value to each health state.

S

Sample

A sample is a subset of the *population* selected according to predefined criteria. The goal of quantitative scientific studies is to draw conclusions about the entire population from which the sample was taken.

SDIL (Soft Drinks Industry Levy)

The SDIL is a tax on *sugary drinks* announced in March 2016 and introduced in April 2018 in the United Kingdom. Manufacturers are subject to a tiered consumption tax on prepackaged sugary drinks, depending on their sugar content: Drinks with less than 5 g of sugar per 100 ml are exempt from the tax, while those with a sugar content between 5 and 8 g per 100 ml are taxed at 0.18£ per liter, and those with more than 8 g per 100 ml are taxed at 0.24£ per liter.

Significance Level

The significance level refers to the predetermined maximum allowable *probability of error* for a statistical test, i.e., the probability with which the test incorrectly indicates a significant result.

Specific Consumption Tax

A specific consumption tax is levied as a fixed monetary amount based on a particular physical characteristic of a product, such as its quantity or sugar content.

Statistical Significance

An event, such as a reduction in sugary drink consumption after the introduction of a tax, is considered statistically significant if the probability of the event occurring by chance is low. In practice, significance is often quantified using the p-value, which indicates the probability of obtaining the observed data, if though the effect is not present in reality. If the p-value is below the predefined *significance level*, the result is considered statistically significant.

Substitution

Substitution (of a product) refers to the replacement of one product with another product.

Sugary Drinks

Sugary drinks (*sugar-sweetened beverages*; short: *SSBs*) are beverages to which sugar (e.g., sucrose, fructose, glucose, corn syrup, honey, malt syrup, etc.) has been added.

T

Tax Pass-Through Rate

The tax pass-through rate indicates the proportion of a tax that is passed on to the final price for consumers.

Tiered Tax

A tiered tax is a form of taxation in which different tax rates are applied within a defined product category that are linked to specific characteristics of the product. An example of this is having different tax rates depending on the amount of sugar content in sugar-sweetened beverages.

U

Uncertainty Interval

An uncertainty interval is a range of possible values for a quantity that takes all types of uncertainties into account, including random and systematic errors, as well as subjective assessments. It differs from the confidence interval in that it reflects not only the statistical uncertainty associated with the randomness of the sample, but also incorporates systematic and subjective factors.

A Modeling Procedures

In modeling studies, a variety of different methodological procedures are used depending on the specific research question and the purpose of the study (Jahn et al., 2022). Modeling procedures that are often used as a basis for decision-making, for example in healthcare, are detailed by Jahn et al. (2022), together with their similarities and differences.

State-Transition Models (Mertens et al., 2022; Siebert et al., 2012) are among the most commonly used modeling methods in healthcare. They model different states (e.g., health states) and transitions between these states (Jahn et al., 2022). The modeling is done either on the basis of cohorts, or at levels of individuals (Siebert et al., 2012). A detailed description of state-transition models can be found in Section A.1. Previous studies investigating the impact of a tax on sugar-sweetened beverages have also used several other analytical methods to examine the relationships between variables. These analytical methods are further explored in Section A.2. Modeling studies often use additional supporting methods for validation or assessment of the results. An explanation of such auxiliary methods can be found in Section A.3.

A.1 State-Transition Models

State-transition modeling is an intuitive, flexible, and transparent approach to computer-based decision-analytic modeling, encompassing various types of simulations (Siebert et al., 2012). A simulation replicates a known or unknown scenario to better understand it. For example, the impacts of the 2004 tsunami in Southeast Asia (a known scenario) or climate change in the future (an unknown scenario) can be investigated (Bungartz et al., 2013).

State-transition models analyze populations either based on cohorts – groups of people who have experienced a certain event at the same time (e.g., same birth year; Statistisches Bundesamt, n. d. b) – or at the individual level, which is referred to as *microsimulation* (Siebert et al., 2012). If interactions between individuals are also included in microsimulation models, it is called *agent-based models* (Arnold et al., 2019). Cohort and microsimulation models do not take into account such interactions between individuals or groups (Siebert et al., 2012), but instead focus more on the development of the population – which is of primary relevance in the context of examining the effects of taxes. Cohort and microsimulations are presented below in more detail.

A.1.1 Cohort Simulation

In a cohort simulation, virtual cohorts of individuals are created to replicate the characteristics and behaviors of a real population (Iskandar, 2018). A cohort refers to a closed group of individuals who share a common characteristic (e.g., birth year or a specific disease) during a given period (Glenn, 2005). These cohorts are selected or assembled to be representative of the overall population being studied, allowing conclusions to be drawn about the entire population (Ethgen & Standaert, 2012). In the simulation, these cohorts are analyzed over a defined time period by simulating various states (e.g., states of health such as disease onset, disease progression, and death; Ethgen and Standaert, 2012; Iskandar, 2018). *Markov chains* are typically used to estimate the probabilities of transitions between different states (Iskandar, 2018; Siebert et al., 2012).

The result of a cohort simulation can provide valuable insights into long-term effects (e.g., costs) of various interventions, thus helping researchers and policymakers assess the potential impacts of different strategies (e.g., health and healthcare system impacts) and allocate resources effectively (Iskandar, 2018). However, cohort models can only provide insights into trends at the cohort level and cannot predict individual behavior (Verified Metrics, n. d.). Additionally, the transition probabilities formulated with Markov chains do not account for events prior to the current state (Siebert et al., 2012). The statistical significance of the results from cohort simulations can also be influenced by various factors, particularly the prevalence of the risk factor or target variable (Brown & Jiang, 2010).

A.1.2 Microsimulation

Microsimulation models capture the behavior of individual units (e.g., people, households, businesses). Compared to macrosimulation models, which focus on entire populations, they offer more specific and detailed analysis possibilities. They are used when heterogeneity within a population or the examination of individual trajectories is important (Spielauer, 2011; Weymeirsch et al., 2024). They also allow the incorporation of past events into future projections (Spielauer, 2011).

The level of detail with which microsimulation models can assess how today's decisions and actions shape the future makes them an attractive tool for policy-making, as they allow the modeling of policy interventions at varying levels of detail. However, microsimulation models typically require significant investments in personnel and hardware due to their complexity (Spielauer, 2011). The accuracy of the results also heavily depends on the underlying assumptions and rules within the model, where even small changes to the input parameters could have significant impact on the results (Emmert-Fees et al., 2024). Another limitation of microsimulation models is the availability of data (Emmert-Fees et al., 2024; Weymeirsch et al., 2024). To achieve accurate and reliable results, microsimulation models require extensive and detailed data at the individual level. Access to sensitive microdata is often highly restricted, and strict data protection regulations must be complied with (Weymeirsch et al., 2024).

A.2 Analytical Methods

To analyze the effects of a tax on sugar-sweetened beverages, previous studies have also used various other analytical techniques to investigate relationships between variables. They are explained below.

A.2.1 Time Series Analysis

A time series analysis is an analysis that looks at how a set of data points evolves over time. The analysis should take into account factors such as seasonality, and trend of the time series (Fahrmeir et al., 2023). Interrupted time series analysis is a special form of time series analysis. It uses the help of piecewise regression models, where the time series interruption represents a targeted intervention (e.g., tax introduction). To analyze changes attributable to this intervention, the data points before and after the intervention are considered.

When investigating and evaluating the effectiveness of measures at the population level where a randomized trial design is not feasible, interrupted time series analysis is considered one of the best alternatives (Kontopantelis et al., 2015). The procedure makes it possible to take into account medium and long term trends that could occur independently of the measure and thus distort the results of a before-and-after comparison (Penfold & Zhang, 2013). A considerable limitation of the (interrupted) time series analysis is that it is difficult to analyze the isolated effects of individual components when they are close together in time. Moreover, the significance of the analysis is only given if the intervention under investigation is the only factor that has changed at the defined time (Mathes et al., 2024). Potentially concurring influences or parallel interventions must therefore be carefully considered and critically discussed. If data is used at the population level, conclusions at the individual level are not possible; this requires time series data that includes measurements from the same individuals (Penfold & Zhang, 2013).

A.2.2 Multivariate Regression

In contrast to univariate regression, multivariate regression analyzes not only one, but several dependent variables at the same time. An example of this is a study that examines the influence of dietary habits such as the consumption of meat, vegetables, grain products, fruit and chocolate on various health parameters such as BMI, blood pressure and cholesterol levels (Backhaus et al., 2023).

One advantage of these regression models is the simultaneous estimation of the model parameters to clarify the relationships between dependent and independent variables. One limitation (as in univariate regression models) is correctly selecting the predictor variables (Johnson & Wichern, 2007). For example, there is a risk of multicollinearity, i.e. a strong correlation of two or more predictors, which leads to difficulties in estimating their individual effects.

A.3 Additional Supporting Procedures

A.3.1 Monte-Carlo-Simulation

Monte Carlo simulations are scenarios in which the same procedure is carried out very often in random constellations, meaning they obey a certain distribution, in order to obtain the desired result at the end by averaging (Bungartz et al., 2013). In the context of simulation studies, this methodology allows for the consideration of uncertainties in the input parameters, which can influence the simulation results. To this end, probability density functions are first estimated for the input parameters, from which samples are then repeatedly drawn. The model is then evaluated for each sample, resulting in sample distributions of the result variables. Relevant outcome statistics are calculated based on these distributions (Harrison, 2010). Monte Carlo simulations are usually used to analyze complex relationships, especially when uncertainties need to be taken into account. However, as this method is based on numerical approximation, it requires large amounts of data and a high computational effort to ensure good results. The quality of the results depends crucially on the validity of the underlying model and the definition of the model parameters.

A.3.2 Sensitivity Analysis

Sensitivity analysis is a method used in various fields, including economics, social sciences, and natural sciences, to study how sensitive a model is to changes in its input parameters. This analysis examines the impact of variations in inputs on a model's output values to determine which parameters have the greatest impact on the results. The aim of the sensitivity analysis is to quantify the uncertainties in the model predictions and to understand how robust the model is to fluctuations in the input values. This analysis helps identify critical parameters that require more accurate or reliable data and helps decision-makers evaluate the reliability and stability of model results (Romeike & Stallinger, 2021).

A.3.3 Difference-in-Differences

The difference-in-differences method is one of the most frequently used methods in studies that evaluate the effectiveness of certain measures and is therefore widely used in business, public policy, health research and other numerous disciplines. It combines before-and-after comparisons with intervention and control group comparisons, making it intuitively understandable and versatile. The method looks at two points in time – before and after the intervention – and compares the difference in the results of the intervention group (with intervention) with the difference in the control group (without intervention). The effect of the intervention results from the difference between the two differences. However, the application can be impaired by problems such as the violation of the parallel trend assumption, heterogeneous effects within the groups, time-dependent variation of the intervention effects, confounding factors or selection biases due to non-random group assignment. Additional adjustments are therefore necessary to minimize biases (Fredriksson & Magalhães de Oliveira, 2019).

A.3.4 Change-in-Change

The change-in-change process is a statistical procedure for evaluating interventions such as political measures. In contrast to the simpler difference-in-differences method, which assumes that the trends between the control and intervention groups would have been the same without intervention, the change-in-change method allows more complex changes in the distribution of the variables to be taken into account. This allows more accurate results to be obtained, especially in cases where the effects differ based on the population. Still however, the process has disadvantages: it is complex and requires extensive calculations and large amounts of data, which makes it difficult to use with limited resources. Modeling errors or biased assumptions can lead to distorted results, which are difficult to interpret, especially in the case of heterogeneous effects (Athey & Imbens, 2006).